This correspondence is being deposited with the United States Postal Service as Express Mail addressed to: Box Patent Application, Assistant Commissioner for Patents, Washington, DC 20231, on March 30, 2001, Express Mail Receipt No. EF055069970US.

MULTI-SERVICE SEGMENTATION AND REASSEMBLY DEVICE

Bidyut Parruck
Chulanur Ramakrishnan

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit under 35 U.S.C. §120 of U.S. Patent Application Serial No. 09/779,381, filed February 7, 2001. The subject matter of U.S. Patent Application Serial No. 09/779,381 is incorporated herein by reference.

TECHNICAL FIELD

This invention relates to multi-service network communications, such as, for example, line card circuitry disposed within routers and switches.

BACKGROUND INFORMATION

Figure 1 (Prior Art) is a diagram of a part of the Internet 1. The Internet, loosely defined, is a collection of networks that are interconnected by devices called "routers". In the illustration, the Internet 1 involves seven networks N1-N7 and five routers R1-R5. A protocol called the Internet Protocol (IP) is used to communicate a message from a source device (a node) on one network to a destination device (a node) on another network. The message is broken up into pieces and each of these pieces is packaged into what is called an "IP packet". These packets may be of varying lengths. The IP packets of the message are then sent from the source to the destination from one network to the next via the routers. The various IP packets can take different paths to get from the source to the destination. When all the IP packets arrive at the destination, they are reassembled to recreate the original message.

This high level IP message can be transported across an individual network using any one of many lower level protocols. Some of the protocols are packet-based

protocols, whereas others of the protocols are cell-based protocols. One packet-based protocol used to transport IP is called Multi-Protocol Label Switching (MPLS). In MPLS, each packet is encapsulated with an MPLS label by the first MPLS device it encounters as it enters an MPLS network. The MPLS device is called an MPLS edge router. The MPLS edge router analyses the contents of the IP header and selects an appropriate MPLS label with which to encapsulate the packet. MPLS packets therefore have varying lengths in the same way that IP packets do. At all the nodes within the network subsequent to the edge router, the MPLS label (and not the IP header) is used to make the forwarding decisions for the packet. Paths through particular nodes in the network are setup from edge to edge, the label defining the particular path its packet will take. Finally, as an MPLS labeled packet leaves the network via an edge router, the edge router removes the MPLS label.

One cell-based lower level protocol used to transport IP over a network is the Asynchronous Transfer Mode (ATM) protocol. In ATM, all packets are of equal length. They are therefore called "cells". A large IP packet is transported over an ATM network by segmenting the large IP packet into a plurality of smaller pieces. Each of the smaller pieces is packaged to become an ATM cell. The ATM cells are then transported across the ATM network. When the ATM cells reach the edge of the ATM network, their payloads are reassembled to reform the large IP packet. In Figure 1, networks N1, N5 and N3 are cell-based ATM networks. Networks N2, N6, N4 and N7 are packet-based MPLS networks.

In the example of Figure 1, networks N3 and N4 are OC-192 high-speed networks adapted to carry traffic over long distances. Router R2 at one end of network N3 may, for example, be located in San Francisco whereas router R4 at the other end of network N3 may be located in New York. Such high-speed long distance networks are often called the "backbone" of the Internet.

In the example of Figure 1, individual users U1-U10 are coupled to the Internet via local area networks. Networks N1, N2 and N7 are local area networks. In one example where the network is a corporate network serving an office building, the users are corporate employees in a building. In an example where the network is a network operated by an Internet Service Provider (ISP), the users are individual customers that pay the ISP to gain access to the Internet.

Consider the situation where users on networks N1 and N2 issue IP messages that are destined to go to destinations to the right side of the backbone such that the messages should go through one of the two back bone networks N3 and N4. In such a case, the IP traffic from networks N1 and N2 is aggregated and supplied to the router access point on the appropriate backbone network. A portion of the Internet called the "Metropolitan Area" performs this function. In the illustration, the metro area includes a router R1 used for aggregating traffic from networks N1 and N2, and for routing that information to the appropriate one of backbone networks N3 and N4.

Figure 2 (Prior Art) is a more detailed view of router R1. Router R1 includes line cards 2-3 for interfacing to ATM networks, other line cards 4 and 5 for interfacing to MPLS networks, and a switch fabric 11. ATM line card 3 is coupled to ATM network N5 such that router R1 can communicate with backbone network N3 via network N5. Similarly, MPLS line card 5 is coupled to MPLS network N6 such that router R1 can communicate with backbone network N4 via network N6. ATM line card 2 is coupled to ATM network N1 via OC-12 fiber optic link 6, SONET multiplexer 7, higher speed OC-48 fiber optic link 8, and SONET multiplexer 9. MPLS line card 4 is coupled to MPLS network N2 via OC-12 fiber optic link 10, SONET multiplexer 7, higher speed OC-48 fiber optic link 8, and SONET multiplexer 9. SONET multiplexer 7 performs time division multiplexing (TDM) to modulate both ATM traffic from network N1 as well as packet MPLS traffic from network N2 onto the same wavelength channel transmitted over the same fiber optic link 8. SONET multiplexer 9 performs the inverse function of time demultiplexing the signal on fiber optic link 8 to extract and separate the ATM traffic from the MPLS traffic.

Router R1, when it receives an IP message from one of networks N1 or N2, determines whether to forward the message on the message's "next hop" to router R2 or R3. In this way IP network information from the users is aggregated in the metro area and is directed to the correct backbone network for communication across long distances to the destination.

A problem may exist if one of the local area networks coupled to router R1 is disconnected or if the type of traffic on that network is changed from MPLS packet traffic to ATM cell traffic or visa versa. Consider the situation in which ATM network

N1 ceases business operations. In that case, the operator of router R1 will likely want to disconnect network N1 from its SONET multiplexer 7 and to couple in the network of another paying customer. For example, the operator may want to disconnect ATM network N1 and to connect in its place MPLS network N7. If this is done, however, MPLS traffic would be received on ATM line card 2. ATM line card 2 is not suitable for coupling to an MPLS network. Consequently, ATM line card 2 may have to be disconnected and a suitable MPLS line card substituted in its place. With the expansion of the Internet and with advances in IP switching technology, it appears that the proportion of packet networks to ATM networks is increasing. Accordingly, as more and more of the networks coupled to a router such as router R1 migrate from one type of traffic to the other, more and more of the line cards of the router will have to replaced. This is undesirable. A solution is desired whereby a smooth and easy migration from one type of traffic to the next is possible without the removal of line cards or the physical manipulation of the router.

Figure 3 is a diagram of one possible approach to the problem involving a line card 12 that handles both ATM and packet traffic. Line card 12 is coupled to a switch fabric of a router by interface 13. Cell and packet traffic received from fiber optic cable 14 and transmitted on fiber optic cable 15 are time division multiplexed/demultiplexed by TDM device 16. Cell traffic is handled by integrated circuit device 17. Packet traffic is handled by integrated circuit device 18. As the relative amounts of cell traffic to packet traffic change, the same line card can be used.

SUMMARY

Although the line card set forth in Figure 3 is satisfactory for some applications, the general approach of Figure 3 involves substantial cost. The non-recurring engineering costs associated with developing an integrated circuit can be quite high. Even though the development of a particular integrated circuit may make technical sense for a given application, it may be economically unreasonable to do so where production volumes of the integrated circuit would be low. For each line card of Figure 3, there is one integrated circuit for handling ATM traffic and one integrated circuit for handling packet traffic. Developing a single integrated circuit having the

functionality of both the ATM device and the packet device might involve less non-recurring engineering costs than developing two different integrated circuits, but the integration would likely result in an undesirably large integrated circuit. Parts of such a single integrated circuit may see little use in certain circumstances. Consider the situation in which the mix of ATM traffic to packet traffic shifts to where there is little or no ATM traffic. The single integrated circuit would involve a data path and associated circuitry for handling ATM traffic that is underutilized or is not used at all. Providing this extra unnecessary circuitry on the single integrated circuit would constitute an unnecessary expense. It would therefore make the line card solution afforded by the single integrated circuit unnecessarily expensive. It would be preferable to get more use out of the circuitry provided on the integrated circuit in order to reduce costs.

Not only might some circuitry be underutilized, but also so might other circuitry become overburdened. In the above situation, for example, more and more processing responsibilities would be required of the packet handling circuitry. If the processing capability of the packet handling circuitry is sized to accommodate a particular load, then increasing the load beyond that load by more and more of the traffic shifting to packet traffic may result in the line card being inadequate to handle the required traffic.

In one novel aspect, the same circuitry on a single Multi-Service Segmentation And Reassembly (MS-SAR) integrated circuit handles both cell traffic and packet traffic. Rather than there being a first data path through the integrated circuit for cell processing, and another data path through the integrated circuit for packet processing, functional blocks along a single data path process cell and/or packet information that passes through the integrated circuit though the single data path. Individual flows passing through the single data path are processed in accordance with the traffic type of the individual flow. Any mix of cell to packet traffic can be accommodated, thereby enabling a smooth migration from one traffic type to another. The MS-SAR can handle dynamic changes in the relative amounts of cell and packet traffic. Production costs associated with the integrated circuit are reduced because the same functional blocks are used, albeit in different ways, to process different types of flows.

In another novel aspect, production volumes may be increased in order to realize economies of scale and to reduce per part cost. To allow production volumes to be increased, the very same MS-SAR integrated circuit is usable not only both as an ingress device and an egress device, but also with either a packet-based switch fabric or a cell-based switch fabric. By providing a single MS-SAR integrated circuit that functions as either an ingress device or an egress device as compared to a device that can function as just one or the other, the number of applications for the MS-SAR integrated circuit is increased. By providing a single MS-SAR integrated circuit that can work with either a packet-based switch fabric or a cell-based switch fabric, the number of applications for the single MS-SAR integrated circuit is increased as compared to the same device that could only function with one type of switch fabric.

In another novel aspect, a single MS-SAR integrated circuit involves a lookup block, a segmentation block, and a reassembly block, wherein traffic flows through a single data path through the lookup block, and then through the segmentation block, and then through the reassembly block. By using two identical such MS-SAR devices, one in an ingress mode and another in an egress mode, information received onto the ingress MS-SAR in either ATM or packet format can be communicated through either a packet-based or a cell-based switch fabric, and can be output from the egress MS-SAR in either ATM or packet format. Information communicated in AAL5 adaptation cells received onto the ingress MS-SAR can be reassembled and output in packet format on the egress MS-SAR. Packet format information received onto the ingress MS-SAR can be segmented and output in the form of AAL5 adaptation cells from the egress MS-SAR. The versatility of the single MS-SAR to handle many different traffic types using the same circuitry further increases the number of applications for which the integrated circuit can be used.

In another novel aspect, individual flows are processed in different ways by an egress MS-SAR. An indication of the type of egress processing to be done by an MS-SAR on a flow is embedded into the flow that is received from a switch fabric onto the egress MS-SAR. The egress MS-SAR reads the embedded indication and performs the type of egress processing indicated. In one embodiment, the indication of the type of egress processing is contained in a switch header (the switch header

can be either a switch header for a cell-based switch fabric or for a packet-based switch fabric), the switch header being added by a first MS-SAR functioning in an ingress mode, the first MS-SAR and the second MS-SAR being substantially identical integrated circuits. In one embodiment, information on how to locate the indication of the type in the information coming into the egress MS-SAR is provided to the egress MS-SAR for each logical input port of the egress MS-SAR. The egress MS-SAR uses: 1) the input port number of the flow, and 2) the information on how to locate the indication for a given input port, to locate the indication in the information coming in from the switch fabric.

In accordance with known ATM adaptation layer protocols, packet data can be transmitted over an ATM network by segmenting the packet into smaller pieces and then transmitting each smaller piece over the ATM network in the form of an ATM cell. After transmission across the ATM network, the data payloads of the individual ATM cells are recovered and are reassembled into the original packet. This segmentation and reassembly process has been traditionally performed on a line card by reassembling packets as the individual ATM cells are received onto the line card. A reassembly context is maintained for each packet being reassembled. This reassembly context may, for example, include a composite cyclic redundancy check (CRC) value that is modified as each ATM cell carrying a portion of the packet is received and processed. The CRC value calculated on the data portions received is checked against a CRC transmitted in a trailer of the last ATM cell to verify that the data carried by the numerous ATM cells has not been corrupted in the segmentation and reassembly process. Once a packet has been checked and reassembled, is it buffered into a payload memory on the line card. If, for example, the line card were to support the simultaneous reassembly of one million packets, then the line card would have to be able to store one million reassembly contexts. This would involve a large amount of memory.

In another novel aspect, packets to be reassembled on the line card in such an adaptation layer process are not reassembled before being buffered on the line card. Rather, the payloads of the individual cells are buffered in payload memory as cells. The line card does not maintain a reassembly context for each such packet being buffered. When the buffered packet information is to be output from the line card,

the cell payloads corresponding to the packet are read out of payload memory and the cell payloads are reassembled to form the packet. In this way, a large number of packets being received onto the line card do not need to be simultaneously reassembled, but rather the number of packets being simultaneously reassembled can be set to have a smaller maximum. In one embodiment, one million flows to be reassembled can be received at one time onto a line card, but only one packet per active line card output port is reassembled at a time. By reducing the maximum number of packets being simultaneously reassembled, the maximum number of reassembly contexts to be stored on the line card is reduced. Reducing the number of reassembly contexts to be stored reduces the amount of memory necessary and thereby reduces line card costs.

Not only is a reassembly context involved the reassembly process, but a segmentation context is also traditionally involved in the segmentation process. Traditionally, packets to be segmented on a line card in accordance with an adaptation layer process are received and stored into payload memory as packets. When a packet is to be output, it is retrieved from payload memory and is segmented into segments such that each of the segments forms the data payload of an ATM cell. To be able to check the integrity of the data when the segments are reassembled, a CRC is calculated on the packet at the time of segmentation and is transmitted in a trailer that is included in the last ATM cell. For each such segmentation process that is going on at the same time, a segmentation context including a partial CRC value is maintained. If a large number such as a million simultaneous output flows is to be supported by the line card, then the large number of segmentation contexts must be stored on the line card. This involves a lot of memory and consequently increases line card cost.

In another novel aspect, packets to be segmented in accordance with an adaptation layer protocol are segmented on a per input port basis as they are received onto the line card and prior to being buffered on the line card. The packets are not buffered on the line card, but rather segments are buffered. Because only one segmentation is performed at a time for a given line card input port, the maximum number of simultaneous segmentations is limited to the number of input ports. By limiting the maximum number of simultaneous segmentations to be

performed, the memory required to store the associated segmentation contexts is reduced. In one embodiment, one million simultaneous flows can be processed, but there are only sixty-four input ports. The amount of memory required for storing segmentation contexts is therefore significantly reduced.

In another novel aspect, an MS-SAR involves a data path such that data received onto the MS-SAR passes through the data path and to a memory manager that stores the data into payload memory. The data is then read out of the payload memory and passes through the remainder of the data path to be output from the MS-SAR. How and when the data is read out of payload memory is controlled by control circuitry. The control circuitry controls the memory manager so that the memory manager retrieves data from payload memory so that it will be output from the MS-SAR in a manner controlled by the control circuitry. In one novel aspect, the MS-SAR is partitioned into two integrated circuits such that the data path circuitry is disposed on one integrated circuit and such that the control circuitry is disposed on another integrated circuit. Partitioning the MS-SAR in this way facilitates future increasing of data throughput rates without redesigning the MS-SAR. To increase data throughput, for example from OC-192 rates to OC-768 rates, multiple data path integrated circuits are disposed in parallel, each of the data path integrated circuits being controlled by the same control integrated circuit. The control integrated circuit has multiple control interfaces, one such interface for coupling to and controlling each of the data path integrated circuits.

In another novel aspect, a router involves a first line card and a second line card. Each of the first and second line cards involves an MS-SAR operating in the ingress mode and an MS-SAR operating in the egress mode. The egress MS-SAR on the second line card can become endangered of being overloaded if, for example, the ingress MS-SAR on the first line card continues to send network information for a flow to the egress MS-SAR on the second line card, but the egress MS-SAR on the second line card is prevented from outputting that information, for example due to congestion at the framer. Consequently, more and more of the network information for the flow will end up having to be buffered by the egress MS-SAR of the second line card. In one novel aspect, the ingress and egress MS-SAR devices of the first line card are linked by a serial bus on the first line card, and the ingress and egress

MS-SAR devices of the second line card are linked by a serial bus on the second line card. If the egress MS-SAR of the second line card is in danger of becoming overloaded, then the egress MS-SAR of the second line card sends an indication of this situation to the ingress MS-SAR of the second line card via the serial bus on the second line card. The ingress MS-SAR of the second line card relays that information to the first line card by outputting a special status switch cell. The special status switch cell is transported across the switch fabric to the egress MS-SAR of the first line card. The egress MS-SAR of the first line card detects the special status switch cell, and relays the indication of the situation to the ingress MS-SAR of the first line card via the serial bus on the first line card. In response to receiving this indication from the serial bus, the ingress MS-SAR on the first line card slows or stops outputting the information that is overburdening the egress MS-SAR on the second line card.

In another novel aspect, an integrated circuit includes a reassembly circuit, wherein in an ingress mode the reassembly circuit uses a flow ID to lookup a switch header in an memory, and wherein in an egress mode the reassembly circuit uses a flow ID to lookup a network protocol header (for example, ATM header or MPLS header) in the memory.

These are but some of the novel aspects. Other novel structures and methods are disclosed in the detailed description below. This summary does not purport to define the invention. The invention is defined by the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 (Prior Art) illustrates a part of the Internet where different traffic types are aggregated.

Figure 2 (Prior Art) is a diagram of a router in the part of the Internet illustrated in Figure 1.

Figure 3 is a diagram illustrative of an approach to solving a problem associated with the network structure of Figure 1. Although the line card as illustrated may and is intended to reflect a product designed by or being designed by Northern Telecom Ltd., adequate details about this line card are not available to the inventors or the

assignee to state here in this patent document that what is shown in Figure 3 is prior art.

Figure 4 is a diagram of a switching device (in this case a router) in accordance with an embodiment of the present invention.

Figure 5 is a diagram of a line card in the router of Figure 4.

Figure 6 is a diagram that sets forth various application types that the router of Figure 4 (involving a pair of MS-SAR devices, one operating in an ingress mode, and the other operating in an egress mode) can carry out.

Figure 7 is a diagram illustrating ingress and egress application types involving a cell-based switch fabric. These application types can be carried out by the router of Figure 4.

Figure 8 is a diagram illustrating ingress and egress application types involving a packet-based switch fabric. These application types can be carried out by the router of Figure 4.

Figure 9 is a diagram of an example in accordance with an embodiment of the present invention wherein a first flow is processed in accordance with ingress application type 3 and egress application type 11, and wherein a second flow is processed in accordance with ingress application type 0 and egress application type 8.

Figure 10 is a diagram illustrating an MS-SAR and associated memories in accordance with an embodiment of the present invention.

Figure 11 is a simplified representation of an MPLS packet of flow #1 in the example of Figure 9.

Figure 12 is a simplified diagram of chunks of flow #1 that are output by the incoming SPI-4 interface block in the example of Figure 9.

Figure 13 is a diagram that illustrates information flow into an ingress MS-SAR in the example of Figure 9.

Figure 14 is a representation of a port table in the lookup engine.

Figure 15A-15C are diagrams of the building of a per flow queue for flow #1 in the example of Figure 9.

Figure 16 is a diagram of a dequeue memory location (stores the header pointer) for one FID.

Figures 17 and 18 are diagrams of the first and second enqueue memory locations (store the tail pointer) for one FID.

Figure 19 is a diagram of a memory location (stores a pointer to a buffer in the body of a queue) for an intermediate BID in a linked list.

Figure 20 is a diagram of an ATM cell of flow #2 in the example of Figure 9.

Figure 21 is a diagram of a 56-byte chunk of flow #2 as output from the incoming SPI-4 interface block of the ingress MS-SAR in the example of Figure 9.

Figure 22 is a diagram of a 64-byte chunk of flow #2 as output from the segmentation block of the ingress MS-SAR in the example of Figure 9.

Figure 23 is a diagram of a per flow queue for flow #2 of the example of Figure 9.

Figure 24 is a diagram of the port calendar in the reassembly block in the MS-SAR.

Figures 25 and 26 are diagrams of the port empty and port full registers in the reassembly block in the MS-SAR.

Figure 27 is a diagram that illustrates information flow from payload memory out of the ingress MS-SAR in the example of Figure 9.

Figure 28 is a diagram that illustrates the three switch cells of flow #1 in the example of Figure 9.

Figure 29 is a diagram of the switch cell for flow #2.

Figure 30 is a diagram that illustrates the general flow of information into egress MS-SAR 200 in the example of Figure 9.

Figure 31 is a diagram of the first switch cell for flow #1 in the example of Figure 9.

Figure 32 is a diagram that illustrates the general flow of information out of egress MS-SAR 200 in the example of Figure 9.

Figure 33 is a diagram that illustrates the format of one FID entry in the header table of the MS-SAR.

Figure 34 illustrates three 64-byte chunks of flow #1 as the chunks pass from reassembly block 205 to outgoing SPI-4 interface block 206 in the example of Figure 9.

Figure 35 illustrates the MPLS packet of flow #1 as output from framer 142 in the example of Figure 9.

Figure 36 illustrates the ATM cell of flow #2 as output from reassembly block 205 in the example of Figure 9.

Figure 37 illustrates the ATM cell of flow #2 as output form outgoing SPI-4 interface block 206 of egress MS-SAR 200 in the example of Figure 9.

Figure 38 illustrates an example of application types 5, 6, 14 and 13.

Figure 39 illustrates the processing of flow #1 in accordance with ingress application type 5 in the example of Figure 38.

Figure 40 illustrates the processing of flow #2 in accordance with ingress application type 6 in the example of Figure 38.

Figure 41 illustrates the processing of flow #1 in accordance with egress application type 14 in the example of Figure 38.

Figure 42 illustrates the processing of flow #2 in accordance with egress application type 13 in the example of Figure 38.

Figure 43 illustrates an example wherein a flow is processed on an ingress MS-SAR in accordance with ingress application type 1 and is processed on an egress MS-SAR in accordance with egress application type 9.

Figure 44 illustrates an example wherein a flow is processed on an ingress MS-SAR in accordance with ingress application type 2 and is processed on an egress MS-SAR in accordance with egress application type 10.

Figure 45 illustrates an example wherein a flow is processed on an ingress MS-SAR in accordance with ingress application type 4 and is processed on an egress MS-SAR in accordance with egress application type 14.

Figure 46 illustrates an example wherein a flow is processed on an ingress MS-SAR in accordance with ingress application type 6 and is processed on an egress MS-SAR in accordance with egress application type 12.

Figure 47 illustrates an example wherein a flow is processed on an ingress MS-SAR in accordance with ingress application type 6 and is processed on an egress MS-SAR in accordance with egress application type 14.

Figure 48 is a diagram of an embodiment wherein MS-SAR functionality is partitioned into two integrated circuit chips (a data path integrated circuit and a control integrated circuit) such that multiple data path integrated circuits chips can be used with one control integrated circuit chip to increase data path throughput.

Figure 49 is a diagram of a packet as it is output from the distribution integrated circuit of Figure 48.

Figures 50 and 51 are diagrams that illustrate the building of a packet queue in connection with the operation of the embodiment of Figure 48.

Figure 52 is a diagram that illustrates a technique for accessing certain information stored in external memory in a reduced amount of time in connection with the embodiment of Figure 48.

Figure 53 is a diagram that illustrates a serial bus that couples an egress MS-SAR of a line card to an ingress MS-SAR of the same line card. The egress MS-SAR can use the serial bus to backpressure the sending ingress MS-SAR.

DETAILED DESCRIPTION

Figure 4 is a simplified diagram of a router 100 in accordance with an embodiment of the present invention. Router 100 includes a plurality of line cards 101-104, a switch fabric 105 and a central processing unit (CPU) 106. The line cards 101-104 are coupled to switch fabric 105 by parallel buses 107-114. In the present example, each of parallel buses 107-114 is a 16-bit SPI-4, Phase II, LVDS parallel bus operating at 400 MHz at a double data rate (DDR). CPU 106 is coupled to line cards 101-104 by another parallel bus 131. In the present example, parallel bus 130 is a 32-bit PCI bus. In this example, each of the line cards can receive network communications in multiple formats. For example, line card 101 is coupled to a fiber optic cable 115 such that line card 101 can receive from cable 115 network communications at OC-192 rates in packets, ATM cells, and/or AAL5 cells. AAL5 cells are considered a type of ATM cell.

Line card 101 is also coupled to a fiber optic cable 116 such that line card 101 can output onto cable 116 network communications at OC-192 rates in packets, ATM cells, and/or AAL5 cells. The fiber optic cables 117 and 118 across which line card 103 communicates are also labeled in the diagram. All the line cards 101-104 in this example have substantially identical circuitry.

Figure 5 is a more detailed diagram of representative line card 101. Line card 101 includes OC-192 optical transceiver modules 119 and 120, two serial-to-parallel devices (SERDES) 121 and 122, a framer integrated circuit 123, a IP classification

engine 124, two multi-service segmentation and reassembly devices (MS-SAR devices) 125 and 126, static random access memories (SRAMs) 127 and 128, and dynamic random access memories (DRAMs) 129 and 130. MS-SAR devices 125 and 126 are identical integrated circuit devices, one of which (MS-SAR 125) is configured to be in an "ingress mode", the other of which (MS-SAR 126) is configured to be in an "egress mode". Each MS-SAR device includes a mode register that is written to by CPU 106 via bus 131. When router 100 is configured, CPU 106 writes to the mode register in each of the MS-SAR devices on each of the line cards so as to configure the MS-SAR devices of the line cards appropriately.

ROUTER AND LINE CARD:

Fiber optic cable 115 of Figure 4 can carry information modulated onto one or more of many different wavelengths (sometimes called "colors"). Each wavelength can be thought of as constituting a different communication channel for the flow of information. Accordingly, optics module 119 converts optical signals modulated onto one of these wavelengths into analog electrical signals. Optics module 119 outputs the analog electrical signals in serial fashion onto a high-speed analog bus 132. Serdes 121 receives this serial information and outputs it in parallel form to framer 123 via high-speed parallel bus 133. Framer 123 receives the information from bus 133, frames it, and outputs it to classification engine 124 via another SPI-4 bus 134. Classification engine 124 performs IP classification and outputs the information to the ingress MS-SAR 125 via another SPI-4 bus 135. The ingress MS-SAR 125 processes the network information in various novel ways (explained below), and outputs the network information to switch fabric 105 (see Fig. 4) via SPI-4 bus 107. All the SPI-4 buses of Figures 4 and 5 are separate SPI-4, phase II, 400 MHz DDR buses having sixteen bit wide data buses.

Switch fabric 105, once it receives the network information, supplies that information to one of the line cards of router 100. Each of the line cards is identified by a "virtual output port" number. Router 100 can include up to 256 line cards. Accordingly, the virtual output port number has a range of from 0 to 255.

In the example of Figure 4, switch fabric 105 may supply network information received on fiber optic cable 115 to any one of fiber optic cables 116, 118, 136 or

137. It is one of the primary functions of router 100 to determine, based on the certain information such as the intended destination of the network information, how to route the information. In the case where the network information is in the IP packet format, the router makes its decision on where to route the network information based on an intended IP destination address present in the IP header of each packet. In the case where the network information is in the ATM cell format, the router makes its decision on where to route the network information based on a virtual path identifier and virtual channel identifier (VPI/VCI) information in the ATM header of each ATM cell.

If, for example, the network information received from fiber optic cable 115 is to be output from fiber optic cable 118, then switch fabric 105 would direct that network information to the "virtual output port" of line card 103. If, on the other hand, the network information received from fiber optic cable 110 is to be output from fiber optic cable 137, then switch fabric 105 would direct that network information to the "virtual output port" of line card 104. To facilitate the rapid forwarding of such network information through the switch fabric 105, network information passed to the switch fabric 105 for routing is provided with a "switch header". The "switch header" may be in a format specific to the manufacturer of the switch fabric of the router. The switch header identifies the "virtual output port" to which the associated network information should be routed. Switch fabric 105 uses the virtual output port number in the switch header to route the network information to the correct line card.

It is, however, generally the router 100 that determines to which of the multiple line cards particular network information will be routed. Accordingly, the router's CPU 106 provisions lookup information in (or accessible to) the ingress MS-SAR 125 so that the MS-SAR 125 will append an appropriate switch header onto the network information before the network information is sent to the switch fabric 105 for routing. The ingress MS-SAR 125 uses header information in the network information to lookup the particular switch header specified by CPU 106 for the particular network information. Once the MS-SAR 125 has found the switch header that it is to append to the network information, MS-SAR 125 appends this switch header to the network information and sends the network information with the switch header on to the switch fabric 105 via SPI-4 bus 107.

Switch fabric 105 receives the network information and forwards it to the line card identified by the particular "virtual output port" in the switch header. Consider the example in which the network information is forwarded to the virtual output port of line card 103 so that the network information from fiber optic cable 115 will be output onto fiber optic cable 117. Because the structures of the various line cards are identical in this example, the flow of network information through line card 103 is explained as passing through the egress MS-SAR 126 of Figure 5. The network information and switch header is received onto the egress MS-SAR 126 of the line card that is identified by the virtual output port number in the switch header. The egress MS-SAR 126 receives the network information, removes the switch header, performs other novel processing (explained below) on the network information, and outputs the network information to framer 123 via SPI-4 bus 138. Framer 123 outputs the network information to serdes 122 via high-speed parallel bus 139. Serdes 122 converts the network information into serial analog form and outputs it to output optics module 120 via high-speed analog bus 140. Output optics module 120 converts the information into optical signals modulated onto one wavelength channel. This optical information is transmitted through fiber optic cable 116 (which corresponds to cable 118).

Although each line card in the example of Figure 4 outputs to only one fiber optic cable (i.e., has only one physical output port), this need not be the case. If, for example, a high speed OC-192 optical transceiver and cable is coupled to a line card, then the line card may only have one physical output port. If, on the other hand, two lower speed optical transceivers (for example, OC-48 transceivers) and cables are coupled to the line card, then the line card may have two physical output ports (one being the first OC-48 transceiver and cable, the other being for the second OC-48 transceiver and cable). The egress MS-SAR 126 outputs onto SPI-4 bus 138 network information to be transmitted along with an output port number from 0 to 63. (This "output port number" is not to be confused with the "virtual output port number" used to identify line cards.) Groups of these 64 output port numbers are mapped by framer 123 to the physical output ports provided on the line card. For example, output ports 0-31 on SPI-4 bus 138 may be mapped to a first physical output port (a first OC-48 output transceiver), whereas output ports 32-63 on SPI-4

bus 138 may be mapped to a second physical output port (a second OC-48 output transceiver). In accordance with the SPI-4 bus protocol, there are 64 logical output ports that egress MS-SAR 126 can specify on SPI-4 bus 138.

In a similar way, a line card may receive network information from more than one optical input transceiver module and more than one fiber optic cable. Information coming into a first optical input transceiver module (i.e., a first physical port) would be supplied onto SPI-4 bus 134 by framer 123 with a logical input port number that maps to the first optical input transceiver module onto which the information was received. Similarly, information coming into a second optical input transceiver module (i.e., a second physical port) would be supplied onto SPI-4 bus 134 by framer 123 with a logical input port number that maps to the second optical input transceiver module onto which the information was received. In accordance with the SPI-4 bus protocol, there are 64 logical input ports that framer 123 can specify on SPI-4 bus 134. The logical input port information on SPI-4 bus 134 flows through classification engine 124 to appear on SPI-4 bus 135.

APPLICATION TYPES:

A flow is a sequence of one or more packets or cells transmitted between a selected source and a selected destination, the flow generally representing a single session using a protocol. Each packet or cell in a flow is expected to have the same routing information according to the dictates of the protocol. The specific embodiment of the MS-SAR device described here can process up to one million flows simultaneously. Each of these one million flows is processed in accordance with any one of several "application types". In each application type, the ingress MS-SAR processes the flow in accordance with one of several "ingress application types" and the egress MS-SAR processes the flow in accordance with one of several "egress application types". Figure 6 sets forth, for each application type, the related ingress application type and the related egress application type. Figure 7 sets forth the ingress and egress application types when the switch fabric is a cell-based switch fabric. Figure 8 sets forth the ingress and egress application types when the switch fabric is a packet-based switch fabric.

EXAMPLE OF APPLICATION TYPES 0, 3, 8 AND 11:

Figures 9-37 illustrate an example of how two flows of different traffic types (flow #1 is an MPLS packet, flow #2 is an ATM cell) are received onto a first line card (in this example, line card 101), simultaneously pass through the ingress MS-SAR 125 on the first line card, pass through switch fabric 105 (in this case switch fabric 105 is a cell-based switch fabric), pass onto a second line card (in this example, line card 103), simultaneously pass through an egress MS-SAR 200 on the second line card 103, and are output by the second line card 103. In this example, flow #1 and flow #2 are both communicated to first line card 101 on the same wavelength channel transmitted on the same fiber optic cable 115. Similarly, flow #1 and flow #2 are both communicated from second line card 103 on the same wavelength channel transmitted on the same fiber optic cable 118.

In Figure 9, flow #1 is an MPLS packet flow passing into an ingress line card, where the ingress line card supplies switch cells to the switch fabric. As indicated by the lower left example of Figure 7, this type of flow is identified as ingress application type 3. Ingress MS-SAR 125 therefore performs processing on flow #1 in accordance with ingress application type 3. In Figure 9, flow #2 is an ATM cell flow passing into an ingress line card, where the ingress line card supplies switch cells to the switch fabric. As indicated by the upper left example of Figure 7, this type of flow is identified as ingress application type 0. Ingress MS-SAR 125 therefore performs processing on flow #2 in accordance with ingress application type 0.

Figure 10 is a more detailed diagram of the MS-SAR devices 125 and 200 present on line cards 101 and 103. Each MS-SAR device includes an incoming SPI-4 interface block 201, a lookup engine block 202, a segmentation block 203, a memory manager block 204, a reassembly and header-adding block 205, an outgoing SPI-4 interface block 206, a per flow queue block 207, a data base block 208, a traffic shaper block 209, an output scheduler block 210, and a CPU interface block 211. In one embodiment, the blocks within dashed line 212 are realized on a first integrated circuit and the blocks within dashed line 213 are realized on a second integrated circuit. In the event the functionality of the MS-SAR is broken into two integrated circuits, a second CPU interface block 214 is provided so that CPU 106 can communicate with both integrated circuits. The MS-SAR interfaces to and uses

numerous other external memory integrated circuit devices 215-228 that are disposed on the line card along with the MS-SAR.

In the example of Figure 9, the MPLS packet of flow #1 is received onto input optics module 119 from fiber optic cable 115. The MPLS packet passes through (see Figures 5 and 9) the input optics module 119, through serdes 121, and to multi-service SONET framer 123. Framer 123 is a framer that includes a demapper for separating ATM cells and MPLS packets that are received on the same wavelength via input optics module 119. Framer 123 in one embodiment is a Ganges S19202 STS-192 POS/ATM SONET/SDH Mapper available from Applied Micro Circuits Corporation, 200 Brickstone Square, Andover, MA 01810. Framer 123 outputs the MPLS packet information of flow #1, along with a logical input port number indicative of the physical input module from which the information was received, in 16-bit pieces, onto SPI-4 bus 134. Classification engine 124 performs IP (Internet Protocol) classification in the case where router 100 performs IP lookup. Classification engine 124 may, in one embodiment, be a classification engine available from Fast-Chip Incorporated, 950 Kifer Road, Sunnyvale, CA 94086. In the particular example of Figure 9, the packet traffic (flow #1) comes into the router in MPLS form and exits the router in MPLS form. Consequently, IP classification is not performed in this embodiment.

Figure 11 is a simplified diagram of the MPLS packet 300 of flow #1 as it is received onto ingress MS-SAR 125. MPLS packet 300 contains an MPLS header 301, and a data payload 303. An SPI-4 start delimiter 302 and an SPI-4 end delimiter 304 frame the packet. MPLS header 301 includes the 20-bit MPLS label. The MPLS header 301 is disposed in the packet between the layer two (L2) header and the layer three (L3) IP header. The L2 header was removed by framer 123 such that beginning of the packet as received onto ingress MS-SAR 125 is the MPLS header.

Ingress MS-SAR 125 (see Fig. 10) receives the packet information in multiple sixteen-byte bursts, sixteen bits at a time, via sixteen input terminals 198. When incoming SPI-4 interface block 201 accumulates 64 bytes (block 201 accumulates 64 bytes in ingress mode and up to 80 bytes in egress mode), it sends the 64-byte chunk to lookup block 202 via 64-bit wide data bus 317. In this way incoming SPI-4

interface block 201 breaks the packet up into 64-byte chunks and communicates the chunks to lookup block 202.

Figure 12 illustrates the three 64-byte chunks into which this particular MPLS packet 300 is broken as output from incoming SPI-4 interface block 201 to lookup block 202. The incoming SPI-4 interface block 201 uses the start delimiter 302 and the end delimiter 304 to identify the beginning and ending of the packet 300. The incoming SPI-4 interface block 201 outputs two additional signals (an end-of-packet signal and a start-of-packet signal) along with each of the 64-byte chunks. Figure 12 shows the end-of-packet (EOP) and start-of-packet (SOP) bits for each of the three 64-byte chunks.

Figure 13 shows information flow through ingress MS-SAR 125. CPU 106 has previously placed lookup information into ingress MS-SAR 125 so that the information in each MPLS packet header can be used by lookup block 202 to find: 1) a particular flow ID (FID) for the flow that was specified by CPU 106, and 2) the ingress application type. The ingress application type, once determined, is used by other blocks of the ingress MS-SAR 125 to configure themselves in the appropriate fashion to process the network information appropriately.

Only one traffic type (for example, MPLS packet or ATM cell) is permitted on each logical input port of SPI-4 bus 135. The traffic type is assigned by CPU 106, but there can only be one traffic type assigned at a given time for a given logical input port. A "port table " in lookup block 202 contains, for each of the sixty-four logical input ports, the allowed traffic type as specified by CPU 106. This allowed traffic type information is provisioned in the port table before the flow starts by CPU 106.

Figure 14 is a conceptual diagram that shows the relationship of the information in the "port table" in lookup block 202. For each logical input port there is a traffic type defined. Lookup block 202 uses the incoming logical port number supplied by framer 123 to lookup the traffic type allowed on the logical input port that packet 300 was received on. Each traffic type has a known format. Accordingly, once the traffic type is known, the lookup block can locate in the header of the incoming flow the particular information that will be used to lookup an associated FID found in a "FID

hash table". This "FID hash table" is stored in external memories 215 and 216 (see Fig. 10).

In the present example of Figure 9, packet 300 is an MPLS packet. Lookup block 202 uses the traffic type found in the "port table" to locate the 20-bit MPLS label within the first chunk of the packet. The MPLS label contains information similar to source and destination addressing information. A hash generator in the lookup block 202 uses the 20-bit MPLS label as a "hash key" to generate a "hash index". This "hash index" is used as an address for the "FID hash table". The content of the FID hash table location addressed contains both the FID and the ingress application type. For more detailed information on how this FID lookup is performed in one particular embodiment, see U.S. Patent Application Serial No. 09/779,381, filed February 7, 2001, by Parruck et al. (the subject matter of which is incorporated herein be reference).

The FID and ingress application type, once determined, are passed via 64-bit wide data bus 318 (see Figure 10), one at a time, to segmentation block 203 along with the consecutive 64-byte chunks of data. Segmentation block 203 stores these 64-bytes chunks on a per port basis. Each 64-byte chunk is stored in a buffer in SRAM within segmentation block 203, and a pointer to the buffer is pushed onto a FIFO of pointers for the input port. The FID and traffic type is also pushed onto the FIFO with the pointer. In the present example, there are three 64-byte chunks in flow #1. Segmentation block 203 calculates a cyclic redundancy check (CRC) value for the data portion of the entire packet, and adds a trailer including this CRC such that the trailer is at the end of the last 64-byte chunk. Segmentation block 203 adds any pad that might be necessary between the end of the data of the last 64-byte chunk and the trailer. Memory manager block 204 pops the FIFO of pointers such that segmentation block 203 forwards the 64-byte chunks one at a time to memory manager block 204 via a 128-bit wide bus 319.

Payload memory 217 contains a large number of 64-byte buffers, each buffer being addressed by a buffer identifier (BID). Some of the 64-byte buffers of payload memory 217 may be used and storing data, whereas others may not be storing data (i.e., are "free"). Per flow queue block 207 maintains a linked list of the BIDs of available buffers, called the "free buffer queue". The "free buffer queue" is stored in

external SSRAM 226 (see Figure 10). When memory manager block 204 receives the first 64-byte chunk 305 of data associated with packet 300, memory manager block 204 issues an "enqueue" command via enqueue command line 320 to per flow queue block 207. This constitutes a request for the BID of a free buffer. Per flow queue block 207 pops the free buffer queue to obtain the BID of a free buffer, and forwards that BID to memory manager block 204 via lines 321. Memory manager block 204 then stores the 64-byte chunk 305 (see Fig. 12) of data for packet 300 in the buffer in payload memory 217 identified by the BID. The writing of the 64-byte chunk of data is indicated in Figure 13 by the upward pointing heavy arrow 322 that extends toward payload memory block 217. For additional details on the operation of the memory manager block 204 and how it stores information into payload memory 217, see U.S. Patent Application Serial No. 09/779,381, filed February 7, 2001, by Parruck et al. (the subject matter of which is incorporated herein by reference).

Per flow queue block 207 also maintains a linked list (i.e., a "queue") of the BIDs for the various 64-byte chunks of each flow that are stored in payload memory 217. Such a linked list is called a "per flow queue". Figure 15A illustrates how per flow queue block 207 builds the linked list for flow #1. Each linked list has a head pointer and a tail pointer. A list of the head pointers for the one million possible FIDs is stored in external SRAM 228 (see Figure 10). A list of the tail pointers for the one million possible FIDs is stored in external SRAM 227.

The head pointer for the FID of flow #1 is set to point to the memory location that stores the BID where the first chunk 305 is stored in payload memory 217. Because there is only one chunk (C1) in the linked list for flow #1, the tail pointer is set to point to the same location. This first chunk 305 of flow #1 is the start of packet 300 as identified by an SOP bit.

Next, the second 64-byte chunk 306 is received by the memory manager block 204. The memory manager block 204 again issues an enqueue command to per flow queue block 207, again obtains a BID of a free buffer, and then writes the second chunk 306 into the payload memory at the location identified by the BID. The per flow queue block 207 pops the free buffer queue to remove the now-used BID from the free buffer queue, and adds the BID for the second chunk 306 to the

linked list for the FID of flow #1. As illustrated in Figure 15B, the BID of the second chunk 306 is added to the linked list for the FID of flow #1.

Next, the third 64-byte chunk 307 is received by the memory manager block 204. The same sequence of storing the chunks into payload memory and adding the BID to the linked list of flow #1 is carried out. Figure 15C illustrates the per flow queue for flow #1 including pointers to these three chunks. In the case of chunk 307, this chunk is the last chunk of the packet as indicated by the EOP bit. The linked list for flow#1 is therefore complete.

Figure 16 is a diagram of a memory location in FID dequeue memory 228 (see Figure 10) that stores the head pointer for a flow. Per flow queue block 207 stores in the location, along with the BID head, other information it receives from memory manager block 204 relating to the chunk. This information includes an EOP bit that indicates whether the chunk is the last chunk of a packet, an SOP bit that indicates whether the chunk is the first chunk of a packet, and the ingress application type. Per flow queue block 207 also stores with the head pointer EFCI, CLP and OAM bits. These bits were extracted by lookup block 202 from the header. There is one memory location such as the one illustrated in Figure 16 for each per flow queue. In the example of Figure 15C, chunk C1 is pointed to by a head pointer in one such memory location.

Figures 17 and 18 are diagrams of two memory locations in FID enqueue memory 227 (see Figure 10) that store a tail pointer. Per flow queue block 207 stores in these locations, along with a pointer to the BID tail, other information on the chunk including the output port number from which the data will eventually be output, the size of the per flow queue, and an indication of a time to live (a TTL bit). There is one such pair of memory locations for each per flow queue. In the example of Figure 15C, chunk C3 is pointed to by information in a pair of two such memory locations.

Figure 19 is a diagram of a memory location in SRAM 226 that stores the pointer for a chunk in a queue between the head and tail. Per flow queue block 207 stores in this location a pointer to the next BID in the per flow queue. The location also stores an EOP indication and an SOP indication. In the example of Figure 15C, there is one such memory location for chunk C2.

In the example of Figure 9, a second flow (flow #2) of a different traffic type (ATM cell) is received onto the same line card 101 that flow #1 was. In this example, flow #2 is also received via the same fiber optic cable 115 and is modulated onto the same wavelength channel that flow #1 was. The 53-byte ATM cell of flow #2 passes through the same optics module 119, serdes 121 and framer 123. Framer 123 removes the fifth byte (the "HEC" Header Error Control byte) of the ATM header and places the remaining 52-byte ATM cell onto the 16-bit SPI-4 bus 134. This information passes through classification engine 124 to SPI-4 bus 135 such that the 52-byte ATM cell is received onto the incoming SPI-4 interface block 201 of ingress MS-SAR 125 via sixteen SPI-4 input terminals 198 (see Fig. 10). Incoming SPI-4 interface block 201 receives the 52-byte ATM cell in multiples of 16-byte bursts, 16 bits at a time.

Figure 20 is a simplified diagram of the ATM cell 308 as received onto ingress MS-SAR 125. ATM cell 308 is 52-bytes long. ATM cell 308 includes an ATM header 309 and a data payload 311. An SPI-4 bus start delimiter 310 and end delimiter 312 frame the ATM cell. Incoming SPI-4 interface block 201 receives the ATM cell and supplies it to lookup block 202, sixty-four bits at a time, via 64-bit wide data bus 317.

The internal data path within the MS-SAR from incoming SPI-4 interface block 201, through lookup block 202, and to segmentation block 203 is 64-bits (8-bytes) wide. Incoming SPI-4 interface block 201 therefore adds a 4-byte pad 314 (see Figure 21) to pad the 52-byte ATM cell up to 56-bytes (a multiple of 8 bytes) before sending the ATM cell to lookup block 202.

Figure 21 is a diagram of ATM cell 308 as it is output from incoming SPI-4 interface block 201 to lookup block 202. Because the entire ATM cell 308 is contained in one 64-byte chunk, the SOP and EOP signals output by the incoming SPI-4 interface block 201 indicate both the start of packet (in this case the packet is a cell) and end of packet (in this case the packet is a cell).

Lookup block 202 receives the 56-byte chunk for ATM cell 308, and from the logical input port number looks up the traffic type from the "port table" (see Figure 14). The "port table" indicates that the traffic type is ATM cells. Lookup block 202 uses the traffic type to locate the 12-bit VPI and 16-bit VCI fields in the ATM cell

header. The hash generator of lookup block 202 uses the located VPI and VCI information as a "hash key" to create a "hash index". This "hash index" is then used to lookup the FID and ingress application type in the "FID table" stored in external memories 215 and 216.

Once the FID and the ingress application type for flow #2 are determined, these values are passed from lookup block 202 to segmentation block 203 via 64-bit bus 318 along with the 56-byte chunk. Segmentation block 203 adds an additional 8-byte pad to pad the ATM cell up to 64-bytes, stores the 64-byte chunk into an SRAM (not shown) in segmentation block 203, and pushes a pointer to that chunk onto its FIFO of pointers for that input port. Memory manager block 204 pops the FIFO of pointers such that segmentation block 203 supplies the 64-byte chunk from its SRAM to memory manager block 203.

Figure 22 is a simplified diagram of the 64-byte chunk as output from segmentation block 203 onto 128-bit bus 319. This 64-byte cell contains the 4-byte pad 314 added by the incoming SPI-4 interface block 201 and the additional 8-byte pad 315 added by segmentation block 203. When memory manager block 204 receives the 64-byte chunk, memory manager block 204 issues an enqueue command via command line 320 to per flow queue block 207, obtains in return a BID, and then stores the 64-byte chunk in payload memory 217 at the location identified by the BID. Per flow queue block 207 pops the free buffer queue, thereby removing the now-used BID from the free buffer queue, and adds the BID to a linked list for flow #2.

Figure 23 is a diagram that illustrates the two linked lists. The linked list for flow #1 (FID1) has BIDs for three chunks C1, C2 and C3, whereas the flow for flow #2 (FID2) has BIDs for one chunk C1. The chunk for the ATM cell is indicated as being both the "start of packet" as well as the "end of packet".

Once the linked lists (queues) for flow #1 and flow #2 are formed, the linked lists are popped (i.e., dequeued) in a particular way and order such that their associated chunks in payload memory 217 are output from ingress MS-SAR 125 in a desired fashion, taking into account user-programmable policing/metering parameters. User-programmable parameters may include, for example, burst size, peak cell rate, and sustainable cell rate.

The dequeue process starts with data base block 208 determining an output port to service using a port calendar. This port calendar is located within data base block 208. Once the output port to service is selected, data base block 208 supplies the output port to traffic shaper block 209 and to output scheduler block 210. The traffic shaper block 209 and the output scheduler block 210 supply flow IDs back to data base block 208. Data base block 208 selects one of the two flow IDs to dequeue for that port. Data base block 208 gives priority to traffic shaper output over output scheduler output such that only if the shaper has no FID to output for a given port will the output scheduler be allowed to schedule an FID for that port. Either traffic shaping is performed on an individual flow by traffic shaper block 209, or output scheduling is performed on the flow by output scheduler block 210.

Traffic shaper block 209 performs traffic shaping on a per flow basis for up to one million flows. On a particular flow, either a single leaky bucket shaping scheme is used, a dual leaky bucket shaping scheme is used, or no shaping at all is used. The shaper selects a peak rate or a sustained rate per flow ID depending on an accumulated credit. Up to 1024 different programmable rates can be shaped simultaneously. In one mode, the rate of incoming traffic with a particular flow ID is measured (i.e., metered) and the flow ID of a violator is marked if the measured rate is above a specific threshold programmed by CPU 106 for the flow ID. Up to 1024 different programmable thresholds can be metered simultaneously.

Output scheduler block 210 uses a weighted round-robin scheme to select a quality of service of the selected port. Once the quality of service is selected, a flow ID is selected based on a round-robin scheme.

When data base block 208 receives a flow ID from either traffic shaper block 209 or output scheduler block 210, data base block 208 generates a request to per flow queue block 207 to issue a dequeue command to dequeue the flow ID. Per flow queue block 207 accesses the per flow queue of the flow ID, determines the next BID to dequeue, and outputs the BID in the form of a dequeue command to memory manager block 204.

Per flow queue block 207 is programmable to cause the linked list of one flow to be output multiple times (i.e., multicast), each time with a different flow ID. Multicasting is performed by replicating the linked list of a flow, each replica linked

list having its own flow ID. Per flow queue block 207 is also programmable to cause multiple flows to be combined into one flow (i.e., tunneling). Tunneling is performed by linking the tail of one linked list to the head of another so as to create one large composite linked list, the composite linked list having one flow ID.

For additional structural and operational details of one particular embodiment of traffic shaper block 209, one particular embodiment of output scheduler block 210, one particular embodiment of data base block 208, and one particular embodiment of per flow queue block 207, see U.S. Patent Application Serial No. 09/779,381, filed February 7, 2001, by Parruck et al. (the subject matter of both of which is incorporated herein by reference).

In the present example, the linked list of flow #1 (see Figure 22) is dequeued first. Per flow queue block 207 pops the BID of the first chunk C1 off the FID1 linked list for flow #1 and forwards that BID to memory manager block 204 in a dequeue command. The dequeue command contains the BID of the 64-byte chunk to be read out of payload memory 217, as well as the ingress application type, an EOP bit, an SOP bit, and the output port number (one of 64 logical output ports on SPI-4 bus 107). The dequeue command is sent via dequeue command line 323. The BID is sent via BID lines 321. Per flow queue block 207 adds the now available BID to the free buffer queue in external memory 226.

In response to receiving the dequeue command, memory manager block 204 retrieves the first chunk C1 identified by the BID and outputs that first chunk C1 to reassembly block 205 via 128-bit data bus 324. Memory manager block 205 supplies to reassembly block 205 other control information including the FID of chunk C1, the SOP and EOP bits, the ingress application type being performed on flow #1, and a logical output port ID (PID) identifying the one of the 64 logical output ports on SPI-4 bus 107 to which the chunk will be sent.

Reassembly block 205 uses the ingress application type to determine what type of action it should perform on the associated chunk. In the present example, flow #1 is processed in accordance with ingress application type 3. External memory 218 (see Fig. 10) contains a "header table" 327. For each FID, the CPU 106 has stored beforehand in header table 327 a switch header that reassembly block 205 is to append to the data of the chunk before sending the data on to the switch fabric.

Accordingly, reassembly block 205 uses the FID to lookup in "header table" 327 the "switch header" placed there for this FID by CPU 106. The switch header can be either 8-bytes or 16-bytes depending on the requirements of the switch fabric. As explained above, the "switch header" contains the "virtual port number" of the particular egress line card that the switch fabric 105 wants the switch cell routed to. In the present example of Figure 9, the egress line card is line card 103. Accordingly, CPU 106 has placed the "virtual port number" of line card 103 into the "switch header" associated with flow #1.

Reassembly block 205 also provides a special 4-byte "Azanda header". The Azanda header is embedded in the switch cell as the last four bytes of the switch header. This Azanda header contains information on how the switch cell should be processed by the particular egress MS-SAR that receives the switch cell (in this case MS-SAR 200). The Azanda header includes the egress application type to be used to process the switch cell in the egress MS-SAR. The Azanda header also includes the FID of the cell, an SOP bit, an EOP bit, and quality of service information.

When reassembly block 205 receives the 64-byte chunk from memory manager block 204, it stores the 64-byte chunk into a dual-port DATA_SRAM and stores the "switch header" (including the "Azanda switch header") into a HDR_SRAM. The DATA_SRAM (not shown) and the HDR_SRAM (now shown) are part of the reassembly block 205. A pointer that points to the data in DATA_SRAM and to the header in HDR_SRAM is pushed onto a queue for the particular logical output port. In the present example, where the MS-SAR is operating in the ingress mode, there is only one logical output port (i.e., the output of line card 101), consequently there is only one output port queue. The pointer is therefore placed on this queue. The output port queue is maintained in a Q_FIFO (not shown) that is also part of the reassembly block 205.

Figure 24 is a conceptual diagram of a "port calendar" located in reassembly block 205. Figure 25 is a diagram of a "port empty" register located in reassembly block 205. Figure 26 is a diagram of a "port full" register located in reassembly block 205. Reassembly block 205 uses the port calendar and registers of Figures 24-26 to determine how to pop the output port queues in the Q_FIFO and move the

associated chunks to an OUTPUT_DATA_FIFO. The OUTPUT_DATA_FIFO (not shown) is located within reassembly block 205 and serves as the interface through which data flows to outgoing SPI-4 interface block 206.

Outgoing SPI-4 interface block 206 periodically sends port full status information that is used to load the "port full" register to keep reassembly block 205 updated on how busy the various output logical ports are. Memory manager block 204 periodically sends empty status information that is used to load the "port empty" register to keep reassembly block 205 updated on whether there is any more information coming from the memory manager 205 for each of the various logical output logical ports. In each of the "port empty" and "port full" registers, there is one bit for each of the respective sixty-four logical output ports of ingress MS-SAR 125.

Reassembly block 205 steps through the rows of the port calendar of Figure 24, one by one. For each row, reassembly block 205 reads the port ID (PID) from the last field of the row and examines the bit corresponding to that port ID in each of the "port empty" and "port full" registers. If the bits in these two registers indicate that the port is neither full (i.e., that port has all the data it can handle) nor empty (no data is available for that port), then reassembly block 205 pops the queue in Q_FIFO for that port ID. The popping of the queue causes reassembly block 205 to retrieve one 64-byte data chunk from DATA_SRAM and its associated header from HDR_SRAM. (In ingress mode, all chunks are 64 bytes but in egress mode EOP chunks can be less than 64 bytes). The header is then appended to the data chunk as the complete "switch cell" is pushed onto the OUTPUT_DATA_FIFO. Outgoing SPI-4 interface block 206 pops the OUTPUT_DATA_FIFO, thereby causing the "switch cell" to be supplied to outgoing SPI-4 interface block 206 in 64-bit pieces via 64-bit bus 325 (see Figure 10). The logical output port is provided to outgoing SPI-4 interface block 206 along with the switch cell. Outgoing SPI-4 interface block 206 in turn outputs the switch cell onto SPI-4 bus 107 in 16-bit pieces via sixteen output terminals 326 of ingress MS-SAR 125. The SPI-4 bus 107 includes an SOP delimiter and an EOP delimiter between data bursts on the SPI-4 bus. Data bursts on the SPI-4 bus are transmitted in multiples of 16 bytes.

In this example, MS-SAR 125 is in the ingress mode, so there is only one logical output port. Consequently there is only one entry in the port calendar of Figure 24.

When a row entry is processed having a jump bit of "1", processing of rows returns to the first row. In the example of Figure 24 where there is only one logical output port, there is a "1" in the jump column of the first row entry for port ID number one. Accordingly, the one and only output queue in the Q_FIFO is popped. The circular arrow in Figure 24 illustrates this return in the present example where there is only one logical output port (PID 1).

Figure 27 illustrates information flow for flow #1 out of ingress MS-SAR 125. The heavy upwardly pointing arrow represents the dequeuing operation being performed by per flow queue block 207 including the passing of BIDs and associated parameters to memory manager block 204. The heavy arrow extending from payload memory block 217 to the right represents the flow of 64-byte chunks through reassembly block 205 and to the switch fabric.

Per flow queue 207 continues issuing dequeue commands in the desired order such that the associated various 64-byte chunks of the two flows (see Figure 23) are read out of payload memory 217 and such that corresponding switch cells are generated. The switch cells are output to switch fabric 105 from outgoing SPI-4 interface block 206 via SPI-4 bus 107.

Figure 28 illustrates the three switch cells that correspond to 64-byte chunks C1, C2 and C3 of packet 300. The switch headers of each of the three cells contain the "virtual port number" of line card 103 to which the cells are to be routed. The Azanda header of each switch cell contains the egress application type (egress application type 11) that the egress MS-SAR 200 is to perform on the switch cells of flow #1.

Figure 29 illustrates the switch cell that corresponds to the one and only 64-byte chunk for ATM cell 308. The switch header of this switch cell contains the "virtual port number" of line card 103 to which the cell is to be routed. The Azanda header of the switch cell contains the egress application type (egress application type 8) that the egress MS-SAR 200 is to perform on the switch cell of flow #2.

Switch fabric 105 receives the switch cells from SPI-4 bus 107 and uses the "virtual port number" embedded in the switch headers to route the switch cells to line card 103. The switch cells are received on the SPI-4 bus 110 (see Fig. 9) of line card 103. SPI-4 bus 110 is coupled to the SPI-4 bus input terminals (see input

terminals 198 of Figure 10) of egress MS-SAR device 200 of line card 103. Operation of egress MS-SAR 200 is explained below in connection with Figure 10.

Figure 30 illustrates the general flow of information into egress MS-SAR 200. The first incoming switch cell of packet 300 is read into incoming SPI-4 interface block 201 and is passed to lookup block 202. Lookup block 202 need not, however, perform a lookup using hashing as was done in the ingress mode described above. Rather, because there is an Azanda header embedded in the switch cells, and because that Azanda header contains the FID and the egress application type to be performed by the egress MS-SAR, lookup block 202 merely locates this Azanda header and extracts the embedded information.

Figure 31 illustrates the first switch cell for packet 300 of flow #1. The Azanda header is the last four bytes of the switch header. Although the size of the switch header can differ from switch fabric to switch fabric, all switch cells for a given switch fabric have the same size. CPU 106 writes the size of the switch header for its switch fabric into a "number of bytes of switch header " field in each row of the "port table" (see Figure 14). This number is the same for all rows of the "port table". Lookup block 202 in egress MS-SAR 200 reads the "number of bytes of switch header" field in the row of the port table in the egress MS-SAR 200 for the logical input port from which the switch cells were received. From this number of bytes of the switch cell, lookup engine block 202 locates the last four bytes which are the Azanda header. From the Azanda header, lookup block 202 retrieves the FID and the egress application type to be performed by egress MS-SAR 200. Lookup block 202 removes the switch header and passes the remaining 64-bytes to segmentation block 203 along with the FID and egress application type. The egress application type is supplied along with the 64-byte chunk so that the other blocks that receive the 64-byte chunk will perform the correct type of processing on the associated 64-bytes. The FID in the example of Figure 9 is flow #1. The egress application type in the example of Figure 9 for flow #1 is egress application type 11.

In accordance with egress application type 11, segmentation block 203 does no segmenting per se but rather forwards the 64-byte chunk to memory manager block 204 via the per-port FIFO mechanism described above. Memory manager block 204 issues an enqueue command via enqueue command line 320, and stores the 64-

byte chunk in payload memory 217. Per flow queue block 207 adds the BID to the per flow queue for flow #1. The same process flow occurs for the second and third switch cells associated with flow #1.

Figure 30 represents the storing of 64-byte chunks into payload memory 217 by the heavy upward pointing arrow that extends toward payload memory block 217. The heavy downward pointing arrow that extends toward per flow queue block 207 represents the enqueue command and the building of the three BID linked list for flow #1.

The switch cell for the ATM cell (see Fig. 29) is received onto egress MS-SAR 200 (see Fig. 9) in similar fashion. As in the case for the switch cells for the packet, lookup block 202 uses the logical input port number supplied by incoming SPI-4 interface block 201 to access the appropriate row in the "port table" (see Fig. 14) in the lookup block 202. From the "number of bytes of switch header" field in that row, lookup block 202 locates the "Azanda header" in the incoming switch cell. The Azanda header contains the egress application type and the FID. Lookup block 202 removes the switch header, and forwards the remaining 64-byte chunk to segmentation block 203 along with the FID and egress application type to be performed by egress MS-SAR 200. The 64-byte chunk in this case contains the 4-byte ATM header, the 48-byte ATM payload, and twelve bytes of pad.

In the example of Figure 9, the FID is flow #2. In the example of Figure 9, the egress application type for flow #2 is egress application type 8. In accordance with egress application type 8, segmentation block 203 does not perform segmentation per se but rather passes the 64-byte chunk and the egress application type to memory manager block 204 via the per-port FIFO mechanism described above. Memory manager block 204 stores the 64-byte chunk. Per flow queue block 207 adds the BID for this chunk and parameters associated with the chunk to a per flow queue for flow #2.

Once the 64-byte chunks from the various switch cells are stored in payload memory 217 and the linked lists for flow#1 and flow #2 are built, then the linked lists can be popped (i.e., "dequeued") and the 64-byte chunks output in various orders. For example, output scheduler block 210 can schedule the popping of the queues to control flow to output ports on a flow ID basis.

Figure 32 illustrates the general flow of information out of egress MS-SAR 200 in one scenario when the linked lists for flow #1 and flow #2 are dequeued. As each BID is dequeued from a per flow queue, the associated 64-byte chunk is read from payload memory 217 and is supplied via 128-bite wide data bus 324 to reassembly block 205. As explained above in connection with the ingress mode, reassembly block 205 maintains one reassembly queue for each of the 64 logical output ports. If two flows share the same logical output port, then the entire linked list for one flow must be dequeued before the linked list for the next flow having the same logical output port is dequeued. In the example of Figure 8, flow #1 and flow #2 have different logical output ports. The dequeue command as received by memory manager block 204 from per flow queue block 207 contains the port ID (PID) of the one of the 64 logical output ports of line card 103.

In the present example, the three-chunk linked list for flow #1 is dequeued first, one chunk at a time. Each 64-byte chunk is read into the DATA_SRAM of reassembly block 205 and a pointer to that chunk is stored in the Q_FIFO in the queue for the logical output port of flow #1. The first chunk contains the MPLS packet header. In an MPLS manipulation operation, reassembly block 205 can add (i.e., "push") MPLS labels to, or remove (i.e., "pop") MPLS labels from, the MPLS header found in the first chunk.

Consider the situation in Figure 9 where a first MPLS label for communication across the network coupled to router 100 via fiber optic cable 115 is to be replaced with a second MPLS label for communication across the network coupled to router 100 via fiber optic cable 118. Reassembly block 205 determines the type of MPLS manipulation to perform using the FID to lookup MPLS information that CPU 106 stored previously in the external "header table" 327 (see Figure 10) of the egress MS-SAR 200. For each FID, header table 327 contains MPLS manipulation information.

Figure 33 is a more detailed diagram of an entry for a FID in "header table" 327 of Figure 10. The entry for each FID includes two 72-bit words. The first word contains an eight-bit control word CONTROL_WORD_1, a first 32-bit field that can contain either a first MPLS label or an ATM VPI/VCI header, and a second 32-bit field that can contain a second MPLS label. The second word contains a second eight-bit

control word CONTROL_WORD_2, a first 32-bit field that can contain a switch header, and a second 32-bit field that holds the 32-bit Azanda header. The first two bits of CONTROL_WORD_1 indicate the header type: "00" indicates MPLS type, "10" indicates ATM type, and "01" indicates L2 Ethernet type. The next three bits of CONTROL_WORD_1 indicate the type of MPLS manipulation to perform: "000" indicates replace MPLS label and decrement the 8-bit MPLS time to live (TTL) field, "001" indicates push one MPLS label and decrement the original MPLS TTL, "010" indicates push two MPLS labels and decrement the original MPLS TTL, "011" indicates pop one MPLS label and decrement the original MPLS TTL, and "100" indicates pop two MPLS labels and decrement the original MPLS TTL. The next two bits of CONTROL_WORD_1 indicate the number of valid header words for that FID in the header table 327. The last bit of CONTROL_WORD_1 indicates that L2 stripping and CRC checking is enabled. The first five bits of CONTROL_WORD_2 indicate the number of L2 header bytes to remove if type is L2 Ethernet.

In the example of Figure 9, the entry in header table 327 for flow #1 indicates that the header type is MPLS and that information in the MPLS field (see Figure 33) determines the action. In the present example, the MPLS label is to be replaced. Reassembly block 205 therefore replaces the MPLS label in the first 64-byte chunk for flow #1 with the MPLS label found in an MPLS label field of the first word of the entry for flow #1 in header table 327. Reassembly block 205 also decrements the original TTL number.

As described above in connection with Figures 24-26, reassembly engine 205 cycles through the entries in its port calendar (see Fig. 24). The port calendar is provisioned beforehand by CPU 106 so that there is one entry for the port of flow #1 and another entry for the port of flow #2. If when reassembly block 205 services one of the rows of the port calendar the full and empty registers indicate that outgoing SPI-4 interface block 206 is ready for data for that output port, then the queue for that output port is popped such that one 64-byte chunk stored in DATA_SRAM is combined with a header from HDR_SRAM if appropriate and the result is pushed onto the OUTPUT_ DATA_FIFO within reassembly block 205.

As the various 64-byte chunks of the packet are received from memory manager block 204, reassembly block 205 uses a CRC engine to calculate a CRC value on

the data. The CRC value calculated on only part of the data is called a "partial CRC". The partial CRC value is changed as each additional piece of data passes through the reassembly block 205. When the last of the data (as indicated by the EOP bit) passes through the CRC engine, the CRC value is a "complete CRC". The complete 32-bit CRC is compared with the 32-bit CRC value found in the AAL5 trailer of the last 64-byte chunk of the packet. If the two CRC values match, then the data was not corrupted.

The AAL5 trailer also contains sixteen-bit "length of payload" field that indicates the length of the payload of the packet being reassembled. Reassembly block 205 maintains a partial packet byte count value for the packet. As each 64-byte chunk of the packet is received from memory manager block 204, reassembly block 205 adds the number of bytes received to the partial packet byte count value. After the last 64-byte chunk for the packet has been received and the partial packet byte count value has been added to, reassembly block 205 compares the now complete packet byte count value with the "length of payload" value from the AAL5 trailer. The two should match.

After checking the CRC and the byte count, reassembly block 205 removes the trailer and any padding from the last 64-byte chunk, and pushes the last chunk onto the OUTPUT_ DATA_FIFO.

Outgoing SPI-4 interface block 206 receives the pieces of the packet, piece by piece (each piece being 64-bytes), from OUTPUT_DATA_FIFO and supplies those pieces of the packet of flow #1 to framer 123. As framer 123 is ready to receive another piece of the packet being sent out on a particular output port, outgoing SPI-4 interface block 206 sends it one via 16-bit wide SPI-4 bus 141. Outgoing SPI-4 interface block 206 supplies bursts and start and end delimiters in accordance with SPI-4 protocol to demarcate the packet. Framer 142 receives the packet from SPI-4 bus 141 and uses the start delimiters and the end delimiters to frame the packet. The packet passes through transmit serdes 143 and optics module 144 and onto fiber optic cable 118. In this way, 64-byte chunks of the packet of flow #1 go through reassembly block 205 one by one, the header of the first chunk being modified in accordance with egress application type 11, the trailer on the last chunk being checked and then removed. Accordingly, the term "reassembly" here does not

necessarily require that the entire packet exist in reassembled form in the reassembly block or even on the MS-SAR. Rather the reassembly block outputs the various chunks of the packet sequentially such that the packet is transmitted in reassembled form on the fiber optic cable.

Figure 34 illustrates the three 64-byte chunks of flow #1 as passed through reassembly block 205 to outgoing SPI-4 interface block 206. The chunks are illustrated as 64-byte chunks rather than 64-bit pieces for ease of illustration.

Figure 35 illustrates the MPLS packet of flow #1 as output from SPI-4 interface block 206 to framer 142.

The 64-byte chunk for the ATM cell of flow #2 in this example is dequeued from payload memory 217 in similar fashion. Per flow queue block 207 issues a dequeue command to memory manager block 204 that indicates the BID of the chunk, and the logical output port from which the flow is to be transmitted. Reassembly block 205 uses the FID of flow #2 to access the FID entry in "header table" 327 (see Figure 9). The header type bits of CONTROL_WORD_1 of the FID entry for flow #2 indicates that the header type is ATM. Reassembly block 205 therefore replaces (i.e., "translates") the VPI/VCI field in the ATM header of the chunk in HDR_SRAM with the VPI/VCI information stored in the FID entry for flow #2. Reassembly block 205 may also be programmed by information in header table 327 to change the CLP (Cell Loss Priority bit in the ATM header that indicates a high or low priority for the ATM cell).

Reassembly block 205 receives the 64-byte chunk, stores it into DATA_SRAM, and pushes a pointer to the 64-bytes onto a queue for the logical output port for flow #2. Reassembly block 205 services the ports in its port calendar as explained above, combines the data portion in the DATA_SRAM with the translated ATM header portion in HDR_SRAM, and loads the completed ATM cell of flow #2 into the OUTPUT_DATA_FIFO in reassembly block 205. The last eight bytes of the 64 bytes in DATA_SRAM are pad. Reassembly block 205 removes this 8-byte pad by not reading it into the OUTPUT_DATA_FIFO. Outgoing SPI-4 interface block 205 pops the OUTPUT_DATA_FIFO such that reassembly block 205 supplies the resulting 56-byte chunk to outgoing SPI-4 interface block 205. Figure 36 illustrates

the 56-byte chunk (ATM cell plus four bytes of pad) as output from reassembly block 205.

Outgoing SPI-4 interface block 206 removes the last four bytes of pad when sending the ATM cell out. It does this using byte enables. Outgoing SPI-4 interface block 206 outputs start and end delimiters to demarcate the 52-byte ATM cell in accordance with the SPI-4 bus protocol. The entire 52-byte cell is output in one burst. Figure 37 illustrates the ATM cell of flow #2 as output from outgoing SPI-4 interface block 206 of egress MS-SAR 200 via 16-bit wide SPI-4 bus 141.

Framer 142 receives the ATM cell from SPI-4 bus 141 and uses the start and end delimiters to frame the ATM cell. Framer 142 then supplies the ATM cell to transmit serdes 143. Serdes 143 in turn supplies the ATM cell in analog serial fashion to optics module 144. Optics module 144 outputs the ATM cell onto fiber optic cable 118. In the example of Figure 9, there is only one physical output port (i.e., the fiber optic cable 118). Framer 142 therefore maps the logical output ports for flow#1 and flow #2 on SPI-4 bus 141 to the same serdes 143 for transmission on the same wavelength channel on the same fiber optic cable 118.

EXAMPLE OF APPLICATION TYPES 5, 6, 14 AND 13:

Figure 38 illustrates an example of how two flows of different types (flow #1 is a flow of three AAL5 cells, flow #2 is an MPLS packet) are received onto a first line card (in this example, line card 101), simultaneously pass through ingress MS-SAR 125 on the first line card, pass through switch fabric 105 (in this example switch fabric 105 is a packet-based switch fabric), pass onto a second line card (in this example, line card 103), simultaneously pass through egress MS-SAR 200 on the second line card 103, and are output by second line card 103. Flow #1 exits line card 103 as an MPLS packet, whereas flow #2 exists line card 103 as four AAL5 cells. In this example, flow #1 and flow #2 are both communicated to first line card 101 on the same wavelength channel transmitted on the same fiber optic cable 115. Similarly, flow #1 and flow #2 are both communicated from second line card 103 on the same wavelength channel and on the same fiber optic cable 118.

In Figure 38, flow #1 is a flow of AAL5 cells flow passing into an ingress line card, where the ingress line card supplies a packet to a packet-based switch fabric. As

indicated by the middle left example of Figure 8, this type of flow is identified as ingress application type 5. The various blocks within ingress MS-SAR 125 therefore perform processing on flow #1 in accordance with ingress application type 5. In Figure 38, flow #2 involves an MPLS packet flow passing into an ingress line card, where the ingress line card supplies a packet to a packet-based switch fabric. As indicated by the bottom left example of Figure 8, this type of flow is identified as ingress application type 6. The various blocks within ingress MS-SAR 125 therefore perform processing on flow #2 in accordance with ingress application type 6.

Ingress application type 5 processing on flow #1 is explained in connection with Figures 38, 5, 39 and 10. The first AAL5 cell of flow #1 is received via fiber optic cable 115 (see Figure 5) and passes through optics module 119, serdes 121, framer 123 and classification engine 124. The 52-byte AAL5 cell is received onto ingress MS-SAR 125 in 16-bit pieces via SPI-4 bus 135 via terminals 198 (see Fig. 10). The AAL5 cell includes a 4-byte ATM header and a 48-byte payload. Incoming SPI-4 interface block 201 appends a 4-byte pad to the end of the 52-byte AAL5 cell before forwarding the resulting 56-byte chunk 400 (see Fig. 39) to lookup block 202. Figure 39 illustrates the 56-byte chunk 400 as it is output from incoming SPI-4 interface block 201.

As explained above in connection with the example of Figure 9, lookup block 202 uses the input port type supplied via bus 317 to lookup the associated traffic type in the "port table" of lookup block 202. In this case, the traffic type is "ATM cells". From the traffic type, lookup block 202 locates the VPI and VCI fields in the ATM cell header. Lookup block 202 uses the VPI and VCI information to lookup in the "FID table" (in external memories 215 and 216) the FID and the ingress application type. Lookup block 202 forwards the FID and ingress application type to the other blocks of the MS-SAR. Lookup block 202 removes the 4-byte ATM header and the 4-byte pad, and forwards the remaining 48-byte chunk 401 of data to segmentation block 203 via bus 318. Figure 39 illustrates the 48-byte chunk 401 as it is output from lookup block 202.

Segmentation block 203 adds a 16-byte pad to the 48-byte chunk to form a 64-byte chunk 402. Memory manager block 204 reads the 64-byte chunk 402 from segmentation block 203 and stores 64-byte chunk 402 into a buffer in payload

memory 217 using the enqueue command mechanism set forth above in connection with the example of Figure 9. The per flow queue for flow #1 at this point involves just one buffer. The other two AAL5 cells of flow #1 are processed in similar fashion such that corresponding 48-byte chunks 403 and 404 are queued in the per flow queue for flow #1.

Ingress application type 6 processing on flow #2 is explained in connection with Figures 38, 5, 40 and 10. The MPLS packet flow #2 is received via fiber optic cable 115 and passes through optics module 119, serdes 121, framer 123 and classification engine 124. The MPLS packet is received onto ingress MS-SAR 125 in 16-bit pieces via SPI-4 bus 135 via terminals 198 (see Fig. 10). The MPLS packet includes a four-byte MPLS header and a data payload. To illustrate operation of the ingress MS-SAR, the MPLS packet data payload in this example is 150 bytes.

Incoming SPI-4 interface block 201 (see Fig. 10) forwards the 154-byte packet 405 as 64-bit pieces to lookup block 202. Figure 40 illustrates the 154-byte packet 405 as it is output from incoming SPI-4 interface block 201. As explained above in connection with the example of Figure 9, lookup block 202 uses the input port type supplied via bus 317 to lookup the associated traffic type in the lookup block's "port table". In this case, the traffic type is "MPLS packet". From the traffic type, lookup block 202 locates the 20-bit MPLS label within the MPLS header. Lookup block 202 uses the MPLS label to determine the FID and the ingress application type. Lookup block 202 forwards the FID and ingress application type to the other blocks of the MS-SAR. The packet 406 is forwarded in 64-bit pieces to segmentation block 203 via bus 318. Although all the 64-bit pieces of a packet for a given input port are supplied to segmentation block 203 before any other 64-bit pieces are received for that input port, 64-bit pieces for other input ports may be communicated from lookup block 202 in an interleaved fashion with the 64-bit pieces of packet 406. Figure 40 illustrates the packet 406 as it is output from lookup block 202.

Segmentation block 203 "segments" the incoming packet 406 into 64-byte AAL-5 like chunks. In the present example, there are three such chunks 407-409. The first chunk 407 contains the 4-byte MPLS header. Segmentation block 203 adds an AAL5 trailer 410 to the end of the last chunk 409. Any gap between the end of the 26-byte data payload 411 of the last chunk 409 and the trailer 410 is filled with a pad

412. The term "segmentation" here does not mean that segmentation block 203 in this embodiment necessarily breaks the incoming packet into 64-byte chunks. That has already been done in this embodiment by SPI-4 interface block 201. Rather, the term "segmentation" means that the packet leaves segmentation block 203 properly segmented into chunks in accordance with a segmentation scheme. In this example, segmentation block 203 calculates a cyclic redundancy check (CRC) value on the chunks of the packet and includes that CRC into trailer 410 of the last chunk 409.

Chunks of the packet are output from incoming SPI-4 interface block 201 in 64-byte chunks. Between the various chunks of a packet, other chunks from other flows may pass through the lookup block 202 and to the segmentation block 203. Segmentation block 203 therefore queues the 64-byte chunks 407-409 into a FIFO on a per port basis. The per port FIFO is located within segmentation block 203. As segmentation block 203 receives each chunk from an input port, segmentation block 203 reads a partial CRC for the port from a CRC table (not shown) within segmentation block 203. The partial CRC is modified as a function of the current chunk received, and the new partial CRC is written back into the CRC table entry for the appropriate port. When the EOP of a chunk indicates the chunk is the last chunk of a packet, then the final CRC is calculated and this CRC is inserted into the trailer.

Memory manager block 204 reads the 64-byte chunks from the per port FIFO in segmentation block 203 and stores the 64-byte chunks into buffers in payload memory 217 using the enqueue command mechanism set forth above in connection with the example of Figure 9. The per flow queue for flow #2 in this case involves three buffers.

Once the 64-byte chunks for flow #1 and flow #2 are stored in payload memory 217, their per flow queues can be dequeued so as to carry out traffic shaping and/or output scheduling functions. In the event that the per flow queue for flow #1 is dequeued, 64-byte chunk 402 (see Figure 39) is output from memory manager block 204 via 128-bit bus 324 (see Figure 10). Reassembly block 205 uses the FID of chunk 402 to lookup a switch header 413 in the "header table" 327. In this case, CPU 106 has placed a switch header 413 appropriate for packet-based switch fabric 105 into the location for flow #1 in "header table" 327. Reassembly block 205

removes the 16-byte pad from the first chunk 402, adds the switch packet header 413 to the front to the chunk, and supplies the combination to outgoing SPI-4 interface block 206 using the port calendar mechanism explained above in connection with the example of Figure 9. This outputting to outgoing SPI-4 interface block 206 can occur before all the other 64-byte chunks of the per flow queue are dequeued. In one embodiment of reassembly block 205, either the second 64-byte chunk of a per flow queue or an EOP chunk must be received before the first chunk can be output. The 64-byte chunks 403 and 404 are dequeued and processed by reassembly block 205 in similar fashion, except that no header is added to these chunks. The combination of the switch header 413 and the data portions of the three chunks 402-404 is a switch packet 414.

Figure 39 shows this switch packet 414 as it is output from reassembly block 205 via 64-bit bus 325. The data portions of the three 64-byte chunks 402-404 together form the data payload of switch packet 414. Figure 39 also illustrates switch packet 414 as it is output from outgoing SPI-4 interface block 206.

The 64-byte chunks 407-409 (see Figure 40) of flow #2 are dequeued in similar fashion, the three chunks being processed through the per port queue in reassembly block 205 for the output port of this flow. Reassembly block 205 uses the FID of the first chunk to lookup a switch header 415 in "header table" 327. Reassembly block 205 adds switch header 415 to the front to the first 64-byte chunk. Reassembly block 205 calculates a CRC from the data payload of the three chunks 407-409 and checks to make sure that it matches the CRC found in the trailer 410 of the last chunk 409. After checking the CRC, reassembly block 205 removes the pad 412 and trailer 410.

First chunk 407 is processed and supplied to the outgoing SPI-4 interface block 206 along with switch header 415 before the remaining chunks 408 and 409 of the packet are dequeued. Reassembly block 205 does not store the entire packet in reassembled form on the MS-SAR, but rather processes and outputs the individual 64-byte chunks of the packet one at a time. In this embodiment, the OUTPUT_DATA_FIFO into which reassembly block 205 pushes processed chunks is only 256 bytes in size, a size inadequate to store the entire 160-byte switch packet.

Chunks of flow #2 are output from reassembly block 205 using the port calendar mechanism explained above in connection with the example of Figure 9. The combination of the data portions of the three chunks 407-409 and the added switch header 415 together form switch packet 416. Figure 40 shows switch packet 416 as it is output from reassembly block 205 via 64-bit bus 325. Switch packet 416 is output from outgoing SPI-4 interface block 206 to the packet switch fabric 105 in 16-bit pieces via terminals 326.

Switch packets 414 and 416 for flow #1 and for flow #2 are switched by the packet switch fabric 105 of router 100 such that they are supplied to line card 103. As described above in connection with the example of Figure 9, router 100 uses a "virtual output port" number in the switch header to identify the particular destination line card to which the packets should be routed.

Processing by egress MS-SAR 200 on flow #1 is explained in connection with Figures 41, 5 and 10. Switch packet 414 from packet-based switch fabric 105 is received in 16-bit pieces onto SPI-4 bus 100 via terminals 198. Switch packet 414 includes 8 or 16 bytes of switch header in addition to the packet itself. Figure 41 illustrates switch packet 414 as it is output from incoming SPI-4 interface block 201. Switch packet 414 is passed to lookup block 202 along with a PID indicating the logical input port.

In the example of Figure 38, flow #1 through egress MS-SAR 200 is a flow from a packet-based switch fabric that is output as an MPLS packet. As indicated by the flow to the bottom right of Figure 8, this is egress application type 14. The Azanda header (see Figure 33) is the last four bytes of the switch header. This Azanda header contains both the FID as well as the egress application type to be performed on flow #1 by egress MS-SAR 200. Again, as explained above in connection with Figure 8, only one traffic type is permitted on any one logical input port. Lookup block 202 uses the logical input port number supplied to it via 64-bit bus 317 (see Figure 10) to locate the Azanda header in the information coming from the switch fabric. Lookup block 202 extracts the FID (FID1) and egress application type (egress application type 14). Lookup block 202 removes the switch header 413 and supplies the data payload of the packet via 64-bit bus 318 to segmentation block 203 along with the extracted FID and egress application type.

Figure 41 illustrates the data payload 417 as it is output from lookup block 202. Payload 417, however, is passed to segmentation block 203 in 64-byte chunks. In the example of Figure 38, the 144-byte payload of flow #1 contains enough information for there to be three such chunks. The first chunk 418 contains 64 bytes, the second chunk 419 contains 64 bytes, and the third chunk 420 contains sixteen bytes of data 421. Segmentation block 203 calculates a CRC on the data of the three chunks and adds a trailer 422 so that the trailer is located at the end of the third 64-byte chunk 420. Segmentation block 203 adds any necessary pad between the end of the data 421 of the last chunk and the trailer 422 of the last chunk. The three chunks 418-420 of Figure 41 are queued on a per port basis into an SRAM in segmentation block 203 as explained above in connection with Figure 40 and pointers to the chunks are pushed onto a FIFO in segmentation block 203. Memory manager block 204 pops the FIFO, reads the chunks, and stores the chunks into payload memory 217. Figure 41 illustrates the three 64-byte chunks 418-420 supplied by segmentation block 203 to memory manager block 204. The enqueue command mechanism as outlined above is used such that a per flow queue of BIDs is formed for the chunks of flow #1.

Processing by egress MS-SAR 200 on flow #2 is explained in connection with Figures 42, 5 and 10. The switch packet 416 from packet-based switch fabric 105 is received in 16-bit pieces onto SPI-4 bus 100 via terminals 198. Switch packet 416 includes 8 or 16 bytes of switch header 415, 4 bytes of MPLS header, and 150 bytes of MPLS data payload.

Figure 42 illustrates switch packet 416 as it is output from incoming SPI-4 interface block 201. Switch packet 416 is passed to lookup block 202 along with the logical input port. The lookup block 202 uses the logical input port supplied to it via 64-bit bus 317 to locate the Azanda header in the information coming from the switch fabric. The lookup block extracts the FID and egress application type (egress application type 13). In the case of flow #2, information from a packet-based switch fabric is output from egress MS-SAR 200 as AAL5 cells. This corresponds to egress application type 13 as indicated in the middle right portion of Figure 8. Lookup block 202 removes switch header 415 and supplies the remainder of the packet (MPLS header and MPLS payload) to segmentation block 203 along with the extracted FID

and egress application type. Figure 41 illustrates the MPLS header 423 and MPLS data payload 424 as supplied by lookup block 202 to segmentation block 203 via 64-bit bus 318.

In egress application type 13, segmentation block 203 "segments" the MPLS header 423 and MPLS data payload 424 into four 48-byte chunks 425-428, because each of these four chunks will be the data payload of an AAL5 cell to be output from egress MS-SAR 200. In this example, the fourth 64-byte chunk contains only ten bytes of data 429. Segmentation block 203 adds a trailer 430 so that the trailer 430 is located at the end of the first 48-bytes of the fourth 64-byte chunk. Any necessary pad 431 is added between the end of data 429 and the start of trailer 430. Segmentation block 203 adds a 16-byte pad to the end of each 48-byte chunk to pad each chunk up to the 64-byte size stored in payload memory 217. Figure 42 illustrates the four 64-byte chunks as output from segmentation block 203. The mechanism for supplying the four chunks to memory block 204 involves a queue as explained above. Memory manager block 204 pops the queue, receives the 64-byte chunks from segmentation block 203, and stores the 64-byte chunks into 64-byte buffers using the enqueue command mechanism described above. A per flow queue for flow #2 is created.

The pointers in the per flow queues for flow #1 and flow #2 are then popped off in a desired order using the dequeue command. Here in this example of Figure 38, we will consider the per flow queue of flow #1 being dequeued first. The three 64-byte chunks for flow #1 are received from payload memory 217 by reassembly block 205 and are pushed onto a queue in reassembly block 205 for the intended output port. The pad and trailer of the third 64-byte chunk 420 (see Fig. 41) are removed. Reassembly block 205 also computes a CRC on the combined data of the three chunks and verifies that the CRC in trailer 422 matches the newly computed CRC. For the first 64-byte chunk 418, reassembly block 205 uses the FID to lookup an MPLS header 432 in header table 327 (see Figure 10). The MPLS header 432 was placed in header table 327 by CPU 106 before the lookup operation. Reassembly block 205 performs an MPLS manipulation operation if controlled to do so by control information in the header table.

Figure 41 illustrates a simplified view of MPLS header 432 and the aggregated 144 bytes of data 433 as it is output from reassembly block 205. MPLS header 432 and data 433 together form an MPLS packet 434. Although shown here reassembled as a 144-byte block, the various 64-byte chunks of packet 434 are output from reassembly block 205 one by one, the first 64-byte chunk having the MPLS header appended to it. The chunks are output from reassembly block 205 using the port calendar and queue mechanism set forth above. Outgoing SPI-4 interface block 206 pops the Q_FIFO of reassembly block 205, receives the chunks of the MPLS packet 434 in 64-bit pieces via 64-bit bus 325, and supplies the pieces of the MPLS packet onto terminals 326. The pieces of MPLS packet 434 pass over SPI-4 bus 141 to framer 142, through framer 142, serdes 143, and optics module 144, and onto fiber optic cable 118. It is therefore seen that ATM cell information received onto the line card 101 passes through the packet-based switch fabric and exits line card 103 as an assembled MPLS packet.

The per flow queue for flow #2 is dequeued next in this example (see Figure 42), the 64-byte chunks 425-428 of flow #2 being supplied to the appropriate output port queue in reassembly block 205. In egress application type 13, reassembly block 205 removes the 16-bytes of pad from each 64-byte chunk to recover the 48-bytes of data. Reassembly block 205 uses the FID of a chunk to lookup a 4-byte ATM header 435, adds the ATM header 435 to the 48-byte chunk of data, and adds a 4-byte alignment pad such that each chunk is 56-bytes. The data 429, pad 431, and trailer 430 of the fourth 64-byte chunk 428 form the data portion of the last 56-byte chunk as illustrated in Figure 42. The ATM header 435 is the same for each of the four 56-byte chunks of this flow.

Reassembly block 205 performs ATM translation by using the FID to access the FID entry in "header table" 327 (see Fig. 10). Because the header type in CONTROL_WORD_1 of the FID entry is "ATM", reassembly block 206 replaces the PTI, CLP, EFCI and OAM bits in the ATM header using the incoming EOP, CLP, EFCI and OAM bits from memory manager block 204. The last AAL5 cell (marked EOP) is marked using the PTI field in the ATM header. The packet length is checked against the maximum MTU. Reassembly block 205 checks the length and CRC in AAL5 trailer 430. Figure 42 illustrates the AAL5 cells 436-439 as output

from reassembly block 205 in 64-bit pieces onto 64-bit bus 325. Outgoing SPI-4 interface block 206 pops the queue for the correct output port, retrieves each AAL5 cell, removes the 4-byte alignment pad, and outputs the resulting 52-byte AAL5 cells onto terminals 326 in 16-bit pieces. The AAL5 cells of flow #2 pass over SPI-4 bus 141, through framer 142, through serdes 143, through optics module 144, and onto fiber optic cable 118. It is therefore seen that MPLS packet information received onto the line card 101 passes through the packet-based switch fabric and exits line card 103 in AAL5 cell form.

APPLICATION TYPES 1 AND 9:

Figure 43 illustrates an example of ingress application type 1 and egress application type 9. The incoming data in this example is two AAL5 cells. These two 52-byte AAL5 cells are received onto ingress MS-SAR 125 via terminals 198. The two cells are output from the incoming SPI-4 interface block. The first cell contains 4 bytes of ATM header and 48 bytes of ATM data #1. The first part of the ATM data #1 in this example is a packet header (for example, an IP header). The second cell contains 4 bytes of ATM header and a small amount of data (ATM data #2). An AAL5 trailer is disposed at the end of the 52 bytes of the second cell, and a pad is disposed between the end of the AAL5 data #2 and the beginning of the trailer in accordance with the AAL5 protocol.

Incoming SPI-4 interface block 201 adds four bytes of pad to each cell prior to sending the cells to lookup block 202. The 4-byte pad is added because the data path from the incoming SPI-4 interface block to the segmentation block is 64-bits wide. Segmentation block 203 removes the 4-byte ATM header and adds an additional twelve bytes of pad for a total of sixteen bytes of pad. Memory manager block 204 passes the resulting two 64-byte chunks to payload memory 217 for storage. The two 64-byte chunks are read from payload memory and are passed to reassembly block 205. Reassembly block 205 does not in this ingress application check the CRC in the trailer of the second cell. Reassembly block 205 adds either 8 or 16 bytes of a switch header to each 64-byte chunk, and passes the resulting chunks to outgoing SPI-4 interface block 206 as illustrated. Chunks are, however, processed through reassembly block 205 one at a time. As illustrated, the packet

header in the AAL5 data #1 is passed through the various blocks of the ingress MS-SAR to the outgoing SPI-4 interface block 206. The cells pass out of the ingress MS-SAR 125 as "switch cells", pass through the cell-based switch fabric, and to a line card having an egress MS-SAR. The arrow from the top portion of Figure 43 to the bottom portion of Figure 43 illustrates this passing through the switch fabric.

The two switch cells pass through the incoming SPI-4 interface block and to lookup block 202 of the egress MS-SAR 200 as illustrated. Lookup block 202 removes the switch headers. The resulting two chunks pass through segmentation block 203 and are passed to memory manager block 204 one at a time for storage. Segmentation block 203 in egress application type 9 does not perform segmentation. The two 64-byte chunks are read out of memory manager block 204 and pass to reassembly block 205. Reassembly block 205 maps AAL5 data #1 and AAL5 data #2 into one contiguous data portion. As each of the data portions passes into reassembly block 205, a CRC on the data is computed. The composite CRC is then compared with the CRC in the trailer of the last 64-byte chunk (in this case the second chunk) as that second chunk passes into reassembly block 205. Reassembly block 205 adds an MPLS header and if programmed to do so by control information in the header table, performs an MPLS manipulation operation on the MPLS header. The resulting packet (MPLS header and AAL5 data #1 and AAL5 data #2) is then sent out to the framer via the outgoing SPI-4 interface block 206.

APPLICATION TYPES 2 AND 10:

Figure 44 illustrates an example of ingress application type 2 and egress application type 10. In this example an incoming MPLS packet (MPLS header and 72 bytes of MPLS payload) is received onto the ingress MS-SAR 125 and passes through incoming SPI-4 interface block 201. Lookup block 202 performs a lookup and determines that the ingress application type is ingress application type 2. Segmentation block 203 outputs the packet as two 64-byte chunks such that the first 48 bytes of the first 64-byte chunk includes the MPLS header and a first part (AAL5 data #1) of the data payload. The remainder of the data of the MPLS packet (AAL5 data #2) is segmented into the second 64-byte chunk. Segmentation block 203 calculates a CRC on the 72 bytes of packet data and includes an AAL5 trailer into

the second 64-byte chunk as illustrated so that the trailer is located at the end of the first 48 bytes of the second chunk. Segmentation block 203 appends 16 bytes of pad to the end of each 48-byte chunk so that each chunk is padded up to 64 bytes. The resulting two 64-byte chunks pass out of segmentation block 203 one at a time and are stored in memory manager block 204. The 64-bytes chunks are read out of payload memory and pass to reassembly block 205 one at a time. Reassembly block 205 performs a switch header lookup and adds a switch header to each 64-byte chunk. Reassembly block 205 does not perform any checking of the CRC in the trailer. The resulting switch cells are output via outgoing SPI-4 interface block 206 as illustrated.

The switch cells pass through a switch-based fabric and to an egress line card that includes the egress MS-SAR 200. The switch cells pass through the incoming SPI-4 interface block 201 of the egress MS-SAR 200. Lookup block 202 locates the Azanda header at the end of the switch header and determines the egress application type (egress application type 10) to be performed. Lookup block 202 removes the switch headers and passes the remaining two 64-byte chunks to the memory manager block 204 for storage into payload memory. The 64-byte chunks are read out of payload memory and are passed one at a time to reassembly block 205. Reassembly block 205 performs a header lookup and adds an ATM header to each 64-byte chunk. Reassembly block 205 removes 12 bytes of pad from the end of each chunk as illustrated in Figure 44. The resulting chunks are passed to outgoing SPI-4 interface block 206. Outgoing SPI-4 interface block 206 last the remaining 4-bytes of pad from the end of each chunk and outputs the chunks as ATM cells to the framer of the egress line card.

APPLICATION TYPES 4 AND 14:

Figure 45 illustrates an example of ingress application type 4 and egress application type 14. In this example an incoming ATM cell (four bytes of ATM header and 48 bytes of ATM payload) is received onto the ingress MS-SAR 125 and passes through incoming SPI-4 interface block 201. Lookup block 202 performs a lookup and determines that the ingress application type is ingress application type 4. Lookup block 202 adds four bytes of pad as illustrated in Figure 45 and passes the

resulting 56 bytes to segmentation block 203. Segmentation block 203 generates neither a CRC nor a trailer. Segmentation block 203 adds an additional 8 bytes of pad as illustrated in Figure 45. The resulting 64-byte chunk is stored by memory manager block 204 in payload memory. The 64-byte chunk is read out of payload memory and passes to reassembly block 205. Reassembly block 205 performs a header lookup and adds an 8 or 16-byte switch header to the 64-byte chunk. The resulting switch packet (8 or 16 bytes of switch header, four bytes of ATM header, 48 bytes of ATM data, and 12 bytes of pad) is passed to a packet-based switch fabric via outgoing SPI-4 interface block 206. The switch packet passes through the switch fabric and is received onto the appropriate line card and the ingress MS-SAR 200. Lookup block 202 locates the Azanda header and determines the egress application type (egress application type 14). Lookup block 202 removes the switch header and passes the remaining 64-bytes to segmentation block 203. Segmentation block 203 does not attach a trailer, but merely passes the 64-byte chunk to memory manager block 204 for storage in payload memory. The 64-byte chunk is retrieved from payload memory and is passed to reassembly block 205. Reassembly block 205 performs a header lookup, and appends an MPLS header as illustrated. Reassembly block 205 removes the last eight bytes of pad. The resulting "encapsulated" ATM cell (MPLS header, 4 bytes of ATM header, 48 bytes of ATM data, and 4 bytes of pad) is output as an MPLS packet to the framer of the egress line card via outgoing SPI-4 interface block 206.

APPLICATION TYPES 6 AND 12:

Figure 46 illustrates an example of ingress application type 6 and egress application type 12. In this example an incoming ATM encapsulated cell (four bytes of MPLS header, 4 bytes of ATM header, and 48 bytes of ATM payload) is received onto the ingress MS-SAR 125 and passes through incoming SPI-4 interface block 201. Lookup block 202 performs a lookup and determines that the ingress application type is ingress application type 6. Lookup block 202 adds four bytes of pad as illustrated in Figure 46 and passes the resulting 56-byte chunk to segmentation block 203. Segmentation block 203 pads the chunk up to 64-bytes but generates neither a CRC nor a trailer. The 64-byte chunk is stored by memory

manager block 204 in payload memory. The 64-byte chunk is read out of payload memory and passes to reassembly block 205. Reassembly block 205 performs a header lookup, adds an 8 or 16-byte switch header to the 64-byte chunk, and removes the last 8 bytes of pad. The resulting packet (8 or 16 bytes of switch header, four bytes of MPLS header, four bytes of ATM header, 48 bytes of ATM data, and 4 bytes of pad) is passed to a packet-based switch fabric via outgoing SPI-4 interface block 206. The switch packet passes through the packet-based switch fabric and is received onto the appropriate line card and the egress MS-SAR 200. In the egress MS-SAR, the lookup block 202 locates the Azanda header and determines the egress application type (egress application type 12). Lookup block 202 removes the switch header and MPLS header and passes the remaining 56-bytes (4 bytes of ATM header, 48 bytes of ATM data, and 4 bytes of pad) to segmentation block 203. Segmentation block 203 does not attach a trailer, but adds an additional 8-byte pad to pad the chunk up to 64 bytes. The 64-byte chunk is stored in payload memory, is read out of payload memory, and passes to reassembly block 205. Reassembly block 205 translates the ATM header. For example, if instructed so by control information in the header table, reassembly block 205 replaces the VPI/VCI information in the ATM header with VPI/VCI information retrieved from the header table. Reassembly block 205 removes the last 8 bytes of pad and passes the resulting ATM cell (4 bytes of ATM header, 48 bytes of data, and 4 bytes of pad) to outgoing SPI-4 interface block 206. Outgoing SPI-4 interface block 206 removes the last 4 bytes of pad and passes the resulting 52-byte ATM cell to the framer of the egress line card.

APPLICATION TYPES 6 AND 14:

Figure 47 illustrates an example of ingress application type 6 and egress application type 14. In this example an incoming MPLS packet (four bytes of MPLS header and 160 bytes of packet payload) is received onto the ingress MS-SAR 125 of a first line card and passes through incoming SPI-4 interface block 201. Lookup block 202 performs a lookup operation and determines that the ingress application type is ingress application type 6. Segmentation block 203 outputs the packet information as 64-byte chunks as illustrated in Figure 47. In the example of Figure

47, the 160 bytes of data payload of the packet is segmented so that the 4-byte MPLS header and 60 bytes of data is segmented into the first 64-byte chunk. The next 64 bytes of data is segmented into the next 64-byte chunk. The remaining 36 bytes of data 440 from the 160-byte payload is segmented into the third 64-byte chunk as illustrated. Segmentation block 203 calculates a CRC on the data payload and places the CRC calculated into a trailer 441 that is incorporated as the end of the last 64-byte chunk. Any space between the end of the data 440 and trailer 441 is filled with a pad 442. Segmentation block 203 may be outputting 64-byte chunks of a packet at the same time that it is receiving chunks of the same packet. The 64-byte chunks are stored by memory manager block 204 into payload memory. The 64-byte chunks are read out of payload memory and are supplied to reassembly block 205. Reassembly block 205 maps the data portions of the 64-bytes chunks into 160-bytes of packet data as illustrated. Reassembly block 205 calculates a CRC on the data as the various chunks of data pass into the reassembly block 205. When the data from the last 64-byte chunk is received onto reassembly block 205, reassembly block compared the final CRC with the CRC in the trailer of the last 64-byte chunk. Using the FID of the flow, reassembly block 205 performs a header lookup and attaches a switch header to the front of the data payload. The resulting switch packet (switch header, MPLS header, and packet data) is supplied to outgoing SPI-4 interface block 206 in pieces, one at a time. Although the packet data payload is illustrated in Figure 46 as being reassembled by reassembly block 205, a first part of the switch packet may be flowing out of reassembly block 205 to outgoing SPI-4 interface block 206 at the same time that a second part of the data portion of the switch packet is being received onto reassembly block 205.

The switch packet of Figure 46 passes through a packet-based switch fabric to a second line card having an egress MS-SAR. The arrow in Figure 46 illustrates this flow of information. The switch packet (switch header, MPLS header, and 160 bytes of packet data in this example) is received onto the egress MS-SAR via incoming SPI-4 interface block 201. Lookup block 202 removes the switch header, locates the Azanda header, and determines from the Azanda header the egress application type to be performed (egress application type 14). Segmentation block 203 segments the packet as illustrated in Figure 46 into 64-byte chunks. Segmentation block adds a

trailer to the end of the last 64-byte chunk as explained above. The 64-byte chunks are supplied one by one to the memory manager block 204 and are stored one by one into payload memory. The 64-bytes are read out of payload memory one by one and are supplied, one by one, to reassembly block 205. When reassembly block 205 receives the first 64-byte chunk, reassembly block 205 performs a header lookup using header table 327. If instructed to by control information found in the header table 327, reassembly block 205 performs an MPLS manipulation operation on the MPLS header. For example, reassembly block 205 may replace the MPLS label in the MPLS header with an MPLS retrieved from header table 327. Reassembly block 205 maps the data portions of the three 64-byte chunks into one packet payload as illustrated in Figure 46. As the data portion of the various 64-byte chunks pass through reassembly block 205, reassembly block 205 calculates a CRC on the data such that reassembly block 205 checks the final composite CRC with the CRC found in the last 64-byte chunk. The first part of the resulting MPLS packet can be output from reassembly block 205 at the same time that later 64-byte chunks for the packet are being received onto reassembly block 205. The resulting MPLS packet (MPLS header, and 160 bytes of packet data) is output to the framer of the egress line card via outgoing SPI-4 interface block 206.

REASSEMBLY "ON-THE-FLY":

Where a segmented packet is received on an egress line card in the form of a plurality of individual cells (for example AAL5 cells), and where the data payloads of the individual cells are to be reassembled into the original packet, the data payloads of the individual cells are buffered in payload memory before any reassembly is done. Rather than doing reassembly before the cells are buffered, reassembly in the MS-SAR 200 is done on a per output port basis as the cells for a flow are being output (i.e., on-the-fly) from the egress MS-SAR. A packet being reassembled is not necessarily reassembled in the sense that the entire reassembled packet exists in reassembled form in reassembly block 205. Rather the individual 64-byte chunks making up the packet are processed by reassembly block 205 one at a time. It is, for example, possible that a first part of the packet has already passed out of the

reassembly block 205 before the last part of the packet has been read out of payload memory 204.

In accordance with one embodiment, a "reassembly context" is maintained for each reassembly process going on. A reassembly context may, for example, include a 32-bit partial CRC and 16-bit partial packet byte count. Rather than maintaining such a reassembly context for each of the many flows coming into the egress MS-SAR (it is, for example, possible that each incoming flow would be AAL5 cells requiring reassembly), reassembly is done on a per output port basis after buffering in payload memory 217, thereby reducing the maximum number of reassembly contexts that have to be maintained to one per output port. There is at most one such packet reassembly going on for each of the output ports that is configured and active. Reducing the number of reassembly contexts to be stored reduces the amount of memory necessary and thereby reduces line card costs.

SEGMENTATION "ON-THE-FLY":

A packet received onto an ingress line card may, in accordance with an adaptation layer protocol, require segmentation into a plurality of pieces, where each of the pieces is the data payload of an individual cell. The AAL5 protocol is an example of one such protocol. The segmenting is done so that the individual cells can later be reassembled into the original packet. Rather than storing such a packet into payload memory, retrieving the packet, and then segmenting it as it is sent out to the switch fabric, segmentation in ingress MS-SAR 125 is done on a per input port basis as flows are received (i.e., on-the-fly) onto the ingress MS-SAR. Segmentation, as the term is used here, encompasses the one-by-one processing of segments of a packet such that the processed segments include information required by the segmentation protocol. A segmentation block may, for example, output a processed segment before the last part of the packet has even been received by the segmentation block.

In accordance with one embodiment, a segmentation context is maintained for each segmentation process going on. A segmentation context may, for example, include a 32-bit CRC value and a 16-bit packet byte count value. Because a packet to be segmented is received onto segmentation block 203 in segments, and

because segmentation block 203 processes these segments one by one, the segmentation engine maintains a partial 32-bit CRC value for the packet and a partial packet byte count value. As each segment is processed, segmentation block 203 updates the partial CRC value and partial packet byte count value. After the last segment has been received onto the segmentation block as indicated by the EOP bit of the segment, then the partial CRC value is the complete CRC value for the packet. Similarly, the partial packet byte count value contains the number of bytes in the packet. The 32-bit complete CRC value and the 16-bit packet byte count value are included by segmentation block 203 into the AAL5-like trailer that the segmentation block appends to the last 64-byte chunk for the packet. Rather than maintaining such a segmentation context for each of the many flows going out of the ingress MS-SAR (it is, for example, possible that each outgoing flow is to be in the form of AAL5 cells), this segmentation process is done before buffering on a per input port basis, thereby reducing the maximum number of segmentation contexts that have to be maintained to one per input port. There is at most one such packet segmentation process going on for each of the input ports that is configured and active. Reducing the number of segmentation contexts to be stored reduces the amount of memory necessary and thereby reduces line card costs. If a separate counter or separate CRC engine is provided for each segmentation process, then reducing the number of segmentation processes going on at one time further reduces costs.

MULTIPLE DATA PATH CHIPS TO INCREASE THROUGHPUT RATE:

Figure 48 illustrates a novel aspect wherein MS-SAR functionality is partitioned into a control integrated circuit and a data path integrated circuit such that system throughput can be increased by using multiple data path integrated circuits. This increase in system throughput is accomplished without having to redesign either the data path integrated circuit or the control integrated circuit. The data path integrated circuit and the control integrated circuit therefore make a versatile chip set. Where a lower throughput is required, cost is reduced by using just one data path integrated circuit with one control integrated circuit. Where a higher throughput is required, multiple data path integrated circuits are used with one control integrated circuit.

Accordingly, the very same data path and control integrated circuits are useful for a wide range of throughput applications. As a result, larger production runs of each integrated circuit is likely and an associated reduction in per part cost is expected.

Figure 48 illustrates a line card 500 having a first MS-SAR 501 configured in the ingress mode and a second MS-SAR 502 configured in the egress mode. Network information passing from fiber optic cable 503 to the switch fabric passes through optics module 504, serdes 505, framer/mapper 506 and classification engine 507 before it reaches ingress MS-SAR 501. In an egress path, network information passes from egress MS-SAR 502, through framer/mapper 506, serdes 508 and optics module 509 to fiber optic cable 510. The optics/serdes/framer circuitry of Figure 48 is similar to the corresponding circuitry of Figure 5, except that the circuitry in Figure 48 can receive from and output to fiber optic cables 503 and 510 are higher data throughout rates. For example, where line card 101 of Figure 5 may receive and output at OC-192 line rates (about 10 gigabits per second), line card 500 of Figure 48 may receive and output at OC-768 line rates (about 40 gigabits per second).

How the partitioning of MS-SAR functionality facilitates handling higher data throughput rates is explained in connection with ingress MS-SAR 501 of Figure 48. Ingress MS-SAR 501 includes a distribution integrated circuit 511, a control integrated circuit 512, four data path integrated circuits 513-516, and one aggregation integrated circuit 517. Control integrated circuit 512 includes the circuitry shown in Figure 10 within dashed line 213. Each of the data path integrated circuits 513-516 includes the circuitry shown in Figure 10 within dashed line 212. Although per flow queue block 207 in the embodiment of Figure 10 has only one control interface for sending and receiving enqueue and dequeue commands, the control integrated circuit 512 of the embodiment of Figure 48 has four such interfaces. Accordingly, ingress control integrated circuit 512 of Figure 48 is coupled to data path integrated circuits 513-516 via control buses 518-521, respectively.

Operation of the circuit of Figure 48 is explained in connection with an example of a high-speed stream of multiple packets being received from fiber optic cable 503 at OC-768 rates. The packets pass, one by one, through optics module 504, serdes 505, framer/mapper 506, and classification engine 507 to distribution integrated

circuit 511. Distribution integrated circuit 511 receives the packets from classification engine 507 via an SPI-4-like bus 522. This bus 522 is like an SPI-4 bus, except that it is capable of operating at 40 gigabits per second rates. Each packet is framed by a start delimiter and an end delimiter.

Distribution integrated circuit 511 distributes each incoming packet to one of the four data path integrated circuits 513-516, the particular data path integrated circuit being chosen so as to distribute data path processing load evenly among the multiple data path integrated circuits. Any of many known load-balancing algorithms can be employed. Distribution integrated circuit 511 only outputs complete packets. In addition to distributing load, distribution integrated circuit 511 includes a sequence number along with each packet. In one embodiment, there is a sequence of sequence numbers for each flow. Distribution integrated circuit 511 employs a lookup block like lookup block 202 of Figure 10. Lookup information is provisioned by the CPU beforehand as explained above in connection with lookup block 202 of Figure 10. Distribution integrated circuit 511 receives a packet, identifies the flow, increments a sequence number for that flow, and adds the sequence number to the packet. Each respective packet of a flow therefore carries a sequence number that is one greater than the sequence number of the previous packet in that flow.

Figure 49 is a diagram of a packet as output by distribution integrated circuit 511. The packet includes a data portion 523 and a header portion 524. The packet is framed by a start delimiter 525 and an end delimited 526. The sequence number 527 added by distribution integrated circuit 511 is disposed between the start delimited 525 and the packet header 524.

The various packets, after being marked with sequence numbers, flow into the various data path integrated circuits 513-516. The lookup block of the data path integrated circuit that receives the packet determines the flow ID of the packet as described above in connection with the embodiment of Figure 10. In the embodiment of Figure 48, the lookup block also extracts the packet sequence number. Both the flow ID and the packet sequence number are then supplied to the control integrated circuit 512 via the enqueue mechanism. Control integrated circuit 512 is therefore aware of which packets of which flows have been received onto which data path integrated circuits. In addition to performing the other traffic

management, policing, shaping, and metering functions described above, control integrated circuit 512 issues dequeue commands to ensure that the various packets of a flow are supplied to aggregation integrated circuit 517 in the same order in which they were received onto distribution integrated circuit 511. To do this, control integrated circuit 512 maintains a "packet queue" for each flow.

Figures 50 and 51 illustrate the building of such a packet queue. In Figure 50, the packets of a flow (FID1) were sequence numbered P1, P2, P3 and so forth by distribution integrated circuit 511. Control integrated circuit 512 maintains a head pointer and a tail pointer for this packet queue. The first packets P1, P2, P3 in this flow have already been dequeued and sent through the aggregation integrated circuit 517 to the switch fabric. Consequently they do not appear in the packet queue. The next packet to be dequeued is packet P5. Control integrated circuit 512 will dequeue the various 64-byte chunks of this packet by dequeuing a per flow queue for this packet as explained above in connection with Figure 10. In the example of Figure 50, packet P7 of flow #1 has not been queued in the packet queue. Accordingly, control integrated circuit 512 will not dequeue packet P8, but rather will wait until packet P7 has been queued.

Figure 51 illustrates the packet queue after packet P7 has been queued. Note that packet P7 has been linked into the queue in its proper place before packet P8 such that the dequeueing of the linked list will result in the packets of the flow being dequeued in the proper order. In the example of Figure 51, packet P5 has been dequeued, and a new packet for this flow, packet P10, has been received and queued.

Control integrated circuit 512 dequeues the various packets in accordance with this scheme so that the various packets of each flow are output to aggregation integrated circuit 517 in the correct order. Aggregation integrated circuit 517 combines the packets in the order it receives them into one stream. That one stream is output to the switch fabric via SPI-4-like bus 528. Bus 528 is like an SPI-4 bus, except that it is capable of operating at 40 gigabits per second rates. It is therefore seen that one high throughput rate data path (40 gigabits/second) coming into the line card is processed using four lower throughput rate data path integrated circuits (10 gigabits/second), and that the outputs of the four lower throughput rate

data path integrated circuits are combined to form one higher throughput rate data path (40 gigabits/second) to the switch fabric.

In the embodiment of Figure 48, the incoming data path having the increased throughput rate is being controlled by only one control integrated circuit 512. The associated increase in processing required of control integrated circuit 512 may result in difficulties in accessing external memory.

Figure 52 is a diagram of an external memory device 529 that is coupled to a control integrated circuit. External memory device 529 may, for example, be external memory 225 of Figure 10. External memory device 529 stores two types of information, information #1 and information #2, both of which must be accessed within a particular amount of time related to the rate of incoming information. Where the control integrated circuit is clocked by a clock signal, this particular amount of time can be referred to as eight clock periods. In the example to the left of Figure 52, each piece of information can be accessed in two clock periods. Both pieces of information are stored in the same external memory device requiring one to be accessed before the other. A total of four clock periods is therefore required to access both pieces of information.

If the circuit of Figure 48 is now employed to handle OC-768 line rate information coming in from fiber optic cable 503, then this same information (information #1 and information #2) must be accessed in a smaller amount of time. Consider the example where both information #1 and information #2 must now be accessed within two clock periods. A memory may not be available that has fast enough access times to handle the accessing required within this fewer number of clock periods.

In accordance with one novel aspect, two external memories are used. Information #1 is stored in the first external memory 530 and information #2 is stored in second external memory 531. The two external memories are accessed at the same time in parallel. If external memories 530 and 531 are the same type of external memory 529, then these external memories have access times of two clock periods and both information #1 and information #2 are accessed within the required two clock periods. It is to be understood, of course, that the eight and two in the example are used only to illustrate the technique of accessing memories in parallel to facilitate handling higher data throughput rates. The technique here is not limited

to the numbers in this example. An example of information #1 is cell count information stored in the embodiment of Figure 10 in PFQ STAT memory 225. An example of information #2 is packet count information stored in the embodiment of Figure 10 in PFQ STAT memory 225. These two types of information are, in one embodiment of Figure 48, stored in different external memory devices.

Although the operation of the embodiment of Figure 48 is described in connection with a flow of packets, the same process is followed to handle flows of cells. The embodiment of Figure 48 handles all the application types of Figures 7 and 8. It is also to be understood that the functionality of the circuit of Figure 48 can be rearranged. For example, the distribution chip and the control integrated circuit may be combined onto one integrated circuit such that only one lookup block is required. Alternatively, the lookup function may be performed on the distribution chip such that results of the lookup are forwarded to the control integrated circuit. Various rearrangements of the functionality of Figure 48 are possible without losing the benefit of the novel partitioning of the MS-SAR.

BACKPRESSING USING SERIAL BUS:

A router 600 involves a first line card 601 and a second line card 601. Each of the first and second line cards involves an MS-SAR operating in the ingress mode and an MS-SAR operating in the egress mode. The egress MS-SAR 603 on the second line card can become endangered of being overloaded if, for example, the ingress MS-SAR 604 on the first line card continues to send network information for a flow to the egress MS-SAR 603 on the second line card, but the egress MS-SAR 603 on the second line card is prevented from outputting that information, for example due to congestion at the framer 605. Consequently, more and more of the network information for the flow will end up having to be buffered by the egress MS-SAR 603 of the second line card (buffered in payload memory).

In one novel aspect, the ingress and egress MS-SAR devices 604 and 606 of the first line card 601 are linked by a serial bus 607 on the first line card, and the ingress and egress MS-SAR devices 603 and 608 of the second line card 602 are linked by a serial bus 609 on the second line card. If the egress MS-SAR 603 of the second line card is in danger of becoming overloaded, then the egress MS-SAR 603 of the

second line card sends an indication of this situation to the ingress MS-SAR 608 of the second line card via the serial bus 609 on the second line card. In one embodiment, it is the per flow queue block of the egress MS-SAR 603 that detects that the size of the free buffer queue has decreased to an undesirably low level. The per flow queue block is therefore coupled to the serial bus 609 as illustrated in Figure 53.

The ingress MS-SAR 608 of the second line card receives this indication from serial bus 609 and relays the indication to the first line card 601 by outputting a special status switch cell 611. The special status switch cell 611 is transported across the switch fabric 610 to the egress MS-SAR 606 of the first line card. The egress MS-SAR 606 of the first line card detects the special status switch cell, and relays the indication of the situation to the ingress MS-SAR 604 of the first line card via the serial bus 607 on the first line card. In one embodiment, it is the segmentation block that detects the special status switch cell. The segmentation block is therefore coupled to the serial bus 607 as illustrated in Figure 53.

In response to receiving this indication from serial bus 607, the ingress MS-SAR 604 on the first line card slows or stops its outputting of the information that is overburdening the egress MS-SAR 603 on the second line card. In one embodiment, the output scheduler is able to slow or stop the outputting of information that is overburdening egress MS-SAR 603. Consequently, the serial bus 607 is coupled to the output scheduler block of ingress MS-SAR 604 as illustrated in Figure 53. For additional details on this backpressuring mechanism, see: U.S. Patent Application Serial No. 09/779,381, filed February 7, 2001, by Parruck et al. (the subject matter of which is incorporated herein by reference).

Although the present invention is described in connection with certain specific embodiments for instructional purposes, the present invention is not limited thereto. Although the bus protocol by which the incoming bus interface block communicates off-chip and the protocol by which the outgoing bus interface communicates off-chip are the same (SPI-4) in the above-described embodiments, they need not be the same. Only one or neither of the interface blocks need communicate off-chip in accordance with the SPI-4 protocol. The SPI-4 bus is mentioned just as an example. More than one incoming bus interface block can be provided so that the

MS-SAR can interface to either of two different types of buses. Although the lookup block is disposed in the data path between the incoming bus interface block and the segmentation block, this need not be the case in all embodiments. The lookup block may, for example, analyze data passing from the incoming bus interface to the segmentation block without acting as a buffer in the data path between the incoming bus interface block and the segmentation block. In some embodiments, the segmentation block does buffer an entire packet before segmenting it, rather than receiving the packet in smaller pieces and then processing those smaller pieces one by one. Similarly, the reassembly block in some embodiments does reassemble an entire packet before sending the reassembled packet out to the outgoing bus interface block. Although in the embodiments described above the incoming network data is stored in payload memory in buffers that all have the same size, this need not be the case. Multiple different buffer sizes are employed in some embodiments. The data of packets could, for example, be stored in one size of buffer whereas the data of ATM cells could be stored in another size of buffer. Although the Azanda header is described here as being embedded in the switch header, this is not the case in all embodiments. In one embodiment, the Azanda header is provided in the switch cell after the payload. The Azanda header is intended to provide a general-purpose vehicle for the communication of information from an ingress MS-SAR to an egress MS-SAR. Information other than FID and egress application type can be included in such an Azanda header. The specific format for the Azanda header provided above is provided to illustrate the general concept. The architectural concepts disclosed are not limited to any particular protocol, but rather are generally applicable to other protocols. Other application types can be supported, where the various functional blocks of the MS-SAR operate on a flow in a particular way determined by the application type of the flow. Accordingly, various modifications, adaptations, and combinations of various features of the described embodiments can be practiced without departing from the scope of the invention as set forth in the claims.